AdventureWorks Sales Forecast

A Probabilistic Approach

Executive Report by Yifan CUI

1. Executive Summary

The purpose of this report is to assess Adventure Works' sales performance for a period of twelve months using Aggregate Monte Carlo Simulation and Daily Bootstrapping. The simulation framework will then provide management with an expected value for revenue, along with a confidence interval within which annual revenue is likely to fall. In addition, it will provide the probability of achieving a specific annual growth target. Traditional sales forecasts are single-point estimates, which obscure the uncertainty of the market and provide management with almost no information for making risk-informed decisions.

The two methods complement each other, representing "top-down" and "bottom-up" perspectives. The Monte Carlo simulation employs the statistical properties of the historical annual growth rate: by sampling the growth rate value itself, it produces a distribution of simulated sales values in the future. This approach is easy to explain and interpret and is particularly useful in terms of determining annual strategic objectives. The Bootstrap simulation resamples historical daily sales figures directly and, being distribution-free, preserves the day-to-day fluctuations in the data, which makes it more suitable for operational-level forecasting.

The Monte Carlo model predicts an expected next-year sales value of $61.4M (90% CI: $55.1M–$67.6M) while the Bootstrap model predicts an expected value of $37.1M (90% CI: $34.6M–$39.6M). The gap between the two is not a modeling mistake, as one might be tempted to conclude, but rather a reflection of differing assumptions about the nature of the market: the former assumes that the recent trend is an accurate description of future trends, while the latter assumes that the future will see a smoothed version of the long-term average of the observed data. What this means is that we recommend using the confidence interval of the more conservative model for budgeting, but reserving the more optimistic model's confidence interval as a reference point to be used only if the monthly monitoring mechanisms outlined in this report indicate the Monte Carlo assumptions have been validated. The report also includes such a monitoring methodology.

2. Business Problem

Going beyond the single-point forecast, the annual planning team at AdventureWorks has three important questions regarding the uncertainty surrounding its sales forecast. First, what is the likely range for next year's sales revenue? Second, what is the probability of achieving a specific growth target (e.g., 10% year-

over-year growth)? Third, how do the competing forecasting approaches differ in their underlying assumptions and what methods should be chosen to meet these business needs? The answers to these questions are critical to the company's annual budgeting and inventory and capacity planning processes, as well as for reporting of risk exposure to the board.

The analysis in this report is based on the complete star-schema dataset AdventureWorks_Sales.xlsx, which contains one fact table, Sales_data (recording individual transaction line items), and five dimension tables: Date_data (calendar dimension), Product_data (product and category dimension), Sales Territory_data (regional dimension), Sales Order_data (channel dimension, distinguishing between the Reseller and Internet sales channels), and Reseller_data and Customer_data (reseller and customer dimensions). The core analytical table was constructed by left-joining Sales_data with Date_data, Product_data, Sales Territory_data, and Sales Order_data respectively. The data spans 2017-07-01 through 2020-06-15 and contains 121,253 transaction records in total; the row count before and after the join is identical, confirming no data loss or duplication occurred.

Before modeling, this report performed a data-quality review, which found: 0 invalid date records, 0 negative sales-amount records, and 0 zero-sales-amount records , resulting in overall good quality of data. There is one point to mention that 2,113 rows have a null SalesTerritoryKey after the join. Those transactions could not be matched to the regional dimension table. As this field is used only for regional-level breakdown analysis, it does not affect the aggregation of the sales time series itself and therefore has no material impact on this report's core modeling results.

3. Methodology

3.1 Part 1 — Aggregate Monte Carlo Simulation (YoY Growth)

The Monte Carlo simulation takes a "top-down" modeling approach. Its basic logic is to sum transaction data that is broken out by calendar year, calculate a year-over-year (YoY) growth rate for each calendar year, and model the future annual growth rate as a normal random variable with a mean and standard deviation equal to those of the historical YoY growth rates. A very large random draw is then simulated from the normal distribution and an annual sales figure for the next year is calculated by applying each simulated growth rate to the estimated sales of the latest full year.

Due to the dataset dating only from July 2017 to June 2020, encompassing two complete years and months of data (2018 and 2019), this paper uses month-end trailing twelve months year-over-year (TTM YoY) growth as opposed to simple year-over-year growth to estimate the parameters on the underlying Normal distribution. Through month-end, calculations are made of the trailing twelve month total for the product being analysed with the previous 12-month total at that month, gathered to give a series of 12

growth points from June 2019 to May 2020 which is used to estimate the mean and standard deviation of the underlying distribution.

From these 12 TTM YoY growth observations, a mean (μ) of 43.23% and a standard deviation (σ) of 8.78% were calculated as shown; for a cross-check, the raw annual YoY growth rate 2018 → 2019 is 40.56% and indeed of a similar magnitude to the TTM mean to support the validity of this estimate; the 2019 total sales ($42,895,109.50) were selected and with the random seed set to 42, 10,000 random draws were made from the Normal distribution N(43.23%, 8.78%), each 10,000 samples representing a plausible growth rate in next year; each of the 10,000 random draws was applied directly to the 2019 total sales baseline, producing a total of 10,000 plausible sales outcomes for next year.

It should be noted that, because the TTM windows overlap, the observations are not fully independent, and the standard deviation may therefore be somewhat understated.

3.2 Part 2 — Daily Sales Simulation via Bootstrapping

The Bootstrap simulation constitutes a "bottom-up" simulation approach which does not rely on any parametric distribution assumption but rather resamples from the existing historical daily sales directly to form a stochastic and realistic scenario of future sales. All transaction data was first combined into a daily sales time series and then it was completed to a calendar ranging from the overall period 2017-07-01 to 2020-06-15 (1,081 days in all) to ensure that the time series was continuous and complete. It was then verified systematically that the completed series contains no days without sales (zero-sales days), i.e. that there were transaction records for all days during the overall period.

Using the 1,081 days of historical daily sales data as a sampling pool. We take the sampling as with replacement: independently sampling one value for each of the next 365 days, thus building one simulated year. Doing this 1,000 times (random seed: 43), we generate 1,000 different 365-day paths of daily sales forecast. The key selling point of using repeated resampling simulations is the flexibility: it does not rely on any pre-specified distribution assumption. This approach will automatically hold many patterns in history data, e.g. the differences among weekdays; the seasonal fluctuations; the volatility of business. This approach may be especially useful for the operational level informal planning.

4. Results

4.1 Part 1 Results

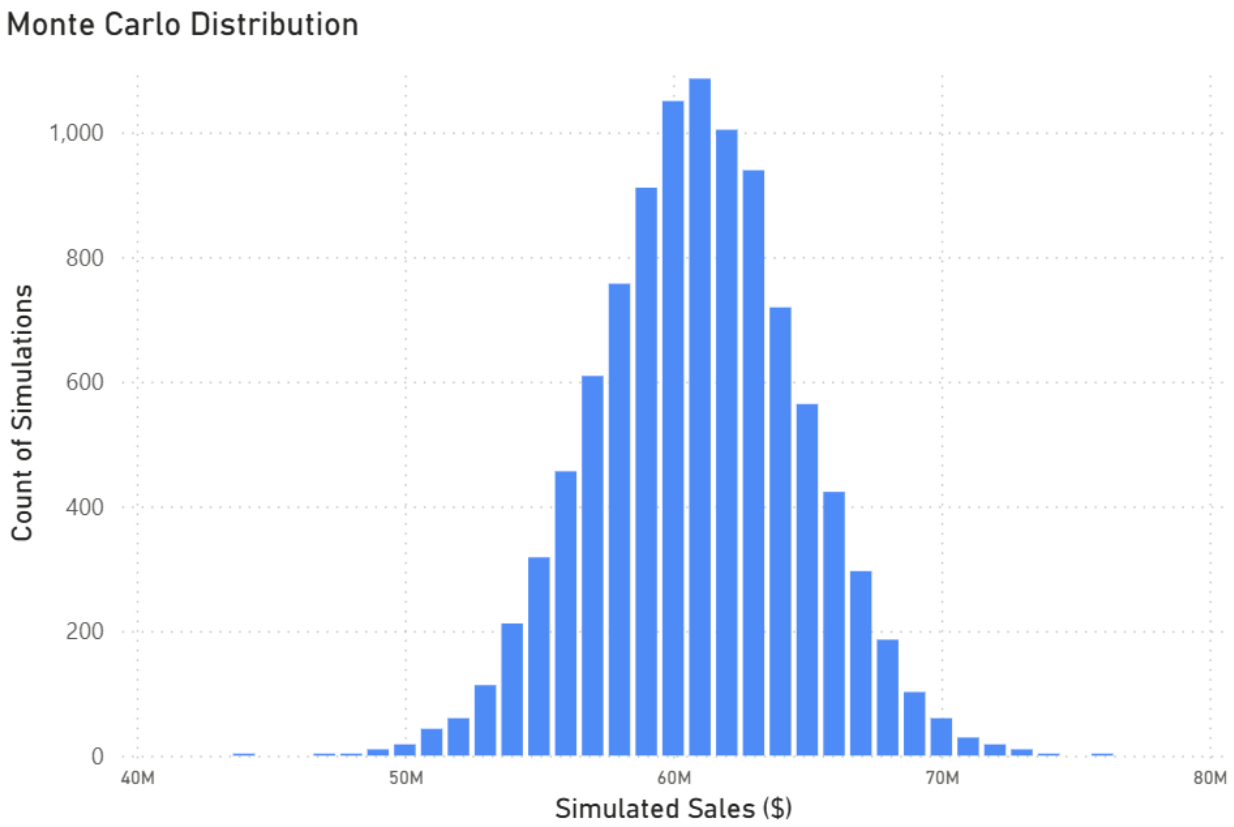


Figure 1. Monte Carlo Annual Sales Forecast (10,000 trials)

Metric	Value
Expected forecast (Mean)	$61,398,421.65
90% CI (5th–95th percentile)	$55,134,671.81 – $67,621,311.21
Probability of exceeding 10% growth vs. 2019	99.99%

The simulation results indicate that the scenario forecast for the next year's sales reflects roughly 43% growth over the 2019 baseline ($42.9M) and 99.99% of simulated scenarios obtain a YoY growth higher than 10%. The former reflects the action of recent upward trend in the TTM YoY growth rate (the mechanism of which is explained in the limitations note in Section 3.1). In general, this model presents a forecast under a "growth momentum continues" scenario that is statistically defensible as a reference point when a company wishes to raise its annual aims or explore upside opportunities.

4.2 Part 2 Results

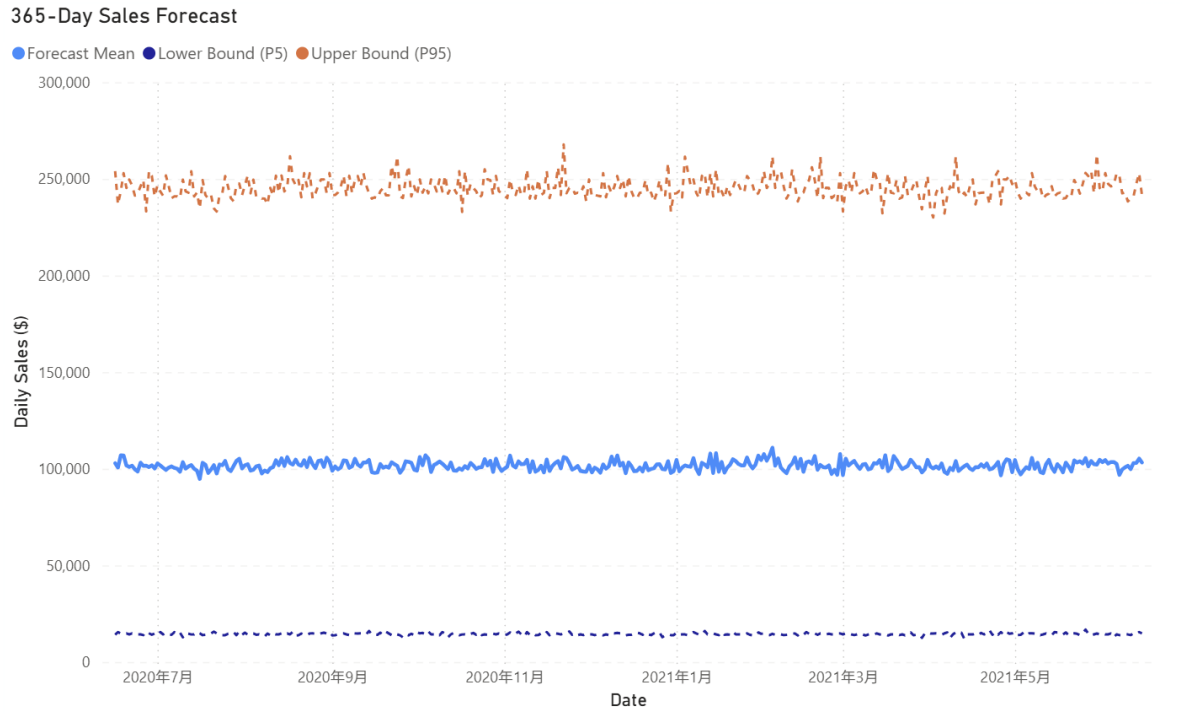


Figure 2. 365-Day Daily Sales Forecast with 90% Confidence Band

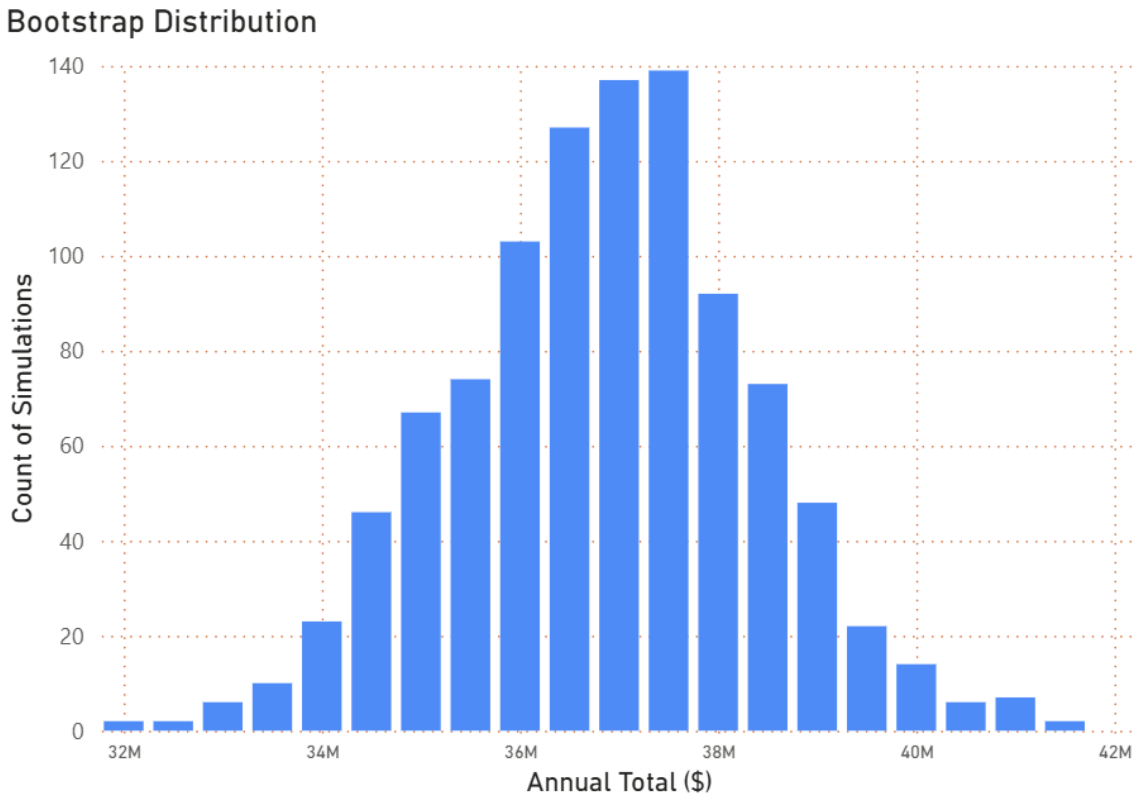


Figure 3. Bootstrap Annual Sales Forecast (1,000 trials)

Metric	Value
Expected forecast (annual total, mean)	$37,103,960.45
90% CI (5th–95th percentile)	$34,633,619.57 – $39,574,655.37

On the daily level, we spot-checked the forecast at Day 1, Day 180 and Day 365. The average at these three points is steady around $100,000~$ 105,000, and the correspondent bounds for the 5th and 95th percentiles are also generally consistent throughout the forecast period. This shows that the model does not extrapolate, or assume continuing growth trend, it fully reflects a sales outlook under the scenario that “the volatility patterns observed historically recur in the future". This property of the Bootstrap model means that it has great value of reference in evaluating a sales baseline under steady-state operating conditions.

5. Comparative Analysis

Metric	Monte Carlo	Bootstrap
Mean	$61,398,421.65	$37,103,960.45
Std Dev	$3,788,479.55	$1,519,351.50
P5	$55,134,671.81	$34,633,619.57
P95	$67,621,311.21	$39,574,655.37
CV (Coefficient of Variation)	6.17%	4.09%

The expected present value of the two models differs by just under $24.3M, a 65.5% relative difference, which is an unacceptably large statistically significant difference. This is not due to a flaw in either's calculation method but to the inherent modeling difference in perspective of "the future" upon which each one operates. In light of its simulation structure, it's obvious that the Monte Carlo model is more extrapolative in nature, because it is centered on the 43.23% mean of the TTM YoY growth observations from June 2019 to May 2020. Because this mean is taken from a historical period of rising growth, the model may be said to implicitly assign it relevance into the future. On the other hand, the Bootstrap model, based on resampling from 1,081 days of historical daily sales, draws each day with a probability equal to its relative frequency in the historical sample, without amplifying recent growth momentum. It is therefore better understood as a historical replay model, assuming that next year will resemble the past. The two models' coefficients of variation are 6.17% and 4.09% respectively, with the Bootstrap model showing lower relative uncertainty. This can be explained by the differing sources of uncertainty in the two models: the Monte Carlo model's uncertainty comes directly from random sampling of the growth rate itself (a standard deviation of 8.78% applied to an already large base sales figure), whereas the Bootstrap model's uncertainty is the result of summing 365 mutually independent daily draws. By the Central Limit Theorem, this summation process substantially compresses the relative volatility of the final annual total.

Collectively, the two models do not represent a "which one is true" relationship; they are simply answering two different and equally important questions about the business. The question that the Monte Carlo model is answering is "how would next year look given that the recently experienced growth momentum may persist into the future?", whereas what the Bootstrap model is answering is "how would next year look if the current market environment resembled market conditions in the last two and half years?" The large divergence between the results of these two models ($37.1M versus $61.4M) also in itself is an indication of significant material risk, which indicates to management that a single forecast number should not simply be used for making an annual plan, but rather development of two separate resource allocation plans based on the two scenarios we have just described ("growth continues" and "authentic past market

conditions") and then allocating the necessary resources to be flexible in case an appropriate adjustment needs to be made at a later date.

Pros & Cons

	Monte Carlo (YoY)	Bootstrap (Daily)
Pros	Simple, highly interpretable modeling logic; well suited to setting annual strategic targets	Fully preserves historical daily volatility patterns, requires no distributional assumption, and provides daily confidence bands
Cons	Relies on a Normal-distribution assumption and struggles to explicitly capture seasonality or autocorrelation	Constrained by the length of the historical observation window; cannot extrapolate to extreme scenarios beyond historical experience
Best suited for	Annual budgeting; board-level strategic reporting	Operational scheduling, inventory management, and cash-flow planning

6. Business Recommendation

Bottom line: The recommendation of this report is for AdventureWorks to use $34.6M to $39.6M (the 90% confidence interval of the Bootstrap model) as the core planning range for next year's sales budget. If the more recent episode of growth acceleration can be confirmed as likely to continue, the planning threshold can be increased to the corresponding range of $55.1M to $67.6M of the Monte Carlo model as reserve reference for an upside scenario; the following gives more detailed application recommendations for different scenarios.

At the level of financial budgeting and board reporting, this report recommends employing the Bootstrap model's 90% confidence interval ($34.6M–$39.6M) to set the lower bound of the annual budget. This model makes no assumption about future growth trends, so its forecast can be seen as the less-controversial and more robust scenario in which "the company's overall operating conditions continue at the past two-and-a-half-year norm," and as the best choice to set a binding financial commitment target , which reduces the operational and reputational risk of missing a budget target.

At a higher level of upside-scenario planning, if the growth acceleration realized in early 2020 (TTM YoY growth accelerated from roughly 35% to more than 50%) is indeed the result of long-term structural changes, for example growth in new product categories and expansion into new markets, rather than a short-term, one-time boost in demand, then the expected confidence level 90% ($55.1M-$67.6M) of the Monte Carlo model can be used as a provisional ceiling for the company to proactively plan for potential future changes in production capacity, inventory stockpiling and channel growth, at an upside scenario. However, we do not advise the company to model the expected value $61.4M as a confining capital target here, as a matter of caution.

To ensure that this distinction translates into business actions, the report also recommends that the company initiate a monitoring mechanism in FY2021 to record the actual TTM YoY growth rate on a month-on-month basis in order to determine whether the growth momentum continues. The company should treat three consecutive months of TTM YoY growth above 45% as confirmation of the "growth momentum continues" scenario and launch the resource expansion plan accordingly. Conversely, if TTM growth falls back to or below the long-term mean (around 35%–40%), this should be treated as a "reversion to historical norm" scenario, in which case the company should hold off on expansion to avoid structural excess in capacity and inventory.

In addition, the report suggests that as more complete annual data is made available over time, the trend behaviour of the TTM YoY growth rate should be periodically re-evaluated in order to better refine the accuracy of the mean and standard-deviation calculated and used to feed the Monte Carlo model developed. Finally, based on the findings of the category- and channel-level breakdown analysis presented

in section 7, the report suggests more differentiated planning approaches across business units. While Bikes (comprising 86% of the total revenue) was the lowest in terms of volatility, it should continue to be the centre of gravity of the overall budget; while Components, Accessories and Clothing (with their higher volatility levels, yet smaller absolute dollar amounts) should instead adopt a more flexible restocking strategy, with less bulk pre-ordering and more short-cycle restocking.

Similarly, at the channel level, Reseller (characterised by bulk B2B orders to the channel) and Internet (with relatively smooth demand patterns) have different inventory-level and cash-flow risk requirements, and hence should have different safety margins. The former should have a more generous inventory and cash-flow safety margin, while the latter can afford to be leaner, in terms of its standing-inventory.

7. Extended Insight — By Category / Channel

To add to the depth of business insight and actionability of the overall final recommendations, this report ran independent Bootstrap simulations in the Category and Channel dimensions along with the main analysis (with 200 simulations and each simulation spanning 365 days; the random seed used was 43 and the same as the main analysis).

By Category (sorted by CV, high to low)

Category	Sales Share	Mean	90% CI	CV
Components	10.75%	$3,988,263.08	$3,543,966.56 – $4,450,581.93	7.19%
Accessories	1.16%	$428,829.80	$379,118.22 – $479,242.25	7.13%
Clothing	1.93%	$715,259.90	$632,937.23 – $792,596.82	6.72%
Bikes	86.17%	$32,026,063.41	$29,753,967.24 – $34,244,182.67	4.14%

By Channel (sorted by CV, high to low)

Channel	Sales Share	Mean	90% CI	CV
Reseller	73.26%	$27,252,184.36	$24,717,961.97 – $29,760,846.58	5.62%
Internet	26.74%	$9,906,231.83	$9,297,661.55 – $10,422,663.17	3.34%

As an internal consistency check, the sum of the four Category means is $37,158,416.19, and the sum of the two Channel means is likewise $37,158,416.19 (the two totals are exactly equal, since both are simply different partitions of the same underlying historical daily sales data). Both figures are also very close to the main Bootstrap model's overall mean of $37,103,960.45 (a difference of about 0.15%, within the range of reasonable error expected from independent segment-level resampling), indicating that the segment-level simulation results are internally consistent with the aggregate-level results.

The simulation results are interpreted by displaying the effect of mixed-signal variability for the top-Level two dimensions of sales: Category and Channel. The Category results follow a fairly typical pattern: the Bikes category accounts for 86.17% of sales and has the highest average daily transaction count, so day-to-day random variation is averaged out across a large number of transactions, giving it the lowest coefficient of variation (4.14%). The Components (10.75%), Accessories (1.16%), and Clothing (1.93%) categories are relatively small in volume with fewer daily transactions and so the contribution of chance (random variation) to total sales on a day-to-day basis relative to total sales is considerably higher; i.e. Component coefficient of variation is slightly higher than Accessories'; though the two values are quite close. Notably, B2B sales (Reseller) account for the largest share of total sales (73.26%), yet these B2B transactions are bulk-ordered and irregular in nature, i.e., less frequent but with a larger individual order value while B2C (Internet channel) is more frequent with smaller individual transaction value. This leads to greater volatility on day-to-day B2B sales due to the chance of a large bulk order closing on a given day. These findings suggest that Mode of transactions (i.e. transaction structure) determine the level of sales volatility in the real world and that inventory planning, staff allocation, and management decisions should be made accordingly.

8. Conclusion

This report has investigated and developed two methods of forecasting the next year's sales for AdventureWorks. The first procedure uses an aggregate Monte Carlo simulation that applies the statistical characteristics of year-on-year growth as gathered from a database of the company's past sales. This simulation provides a range of possible outcomes for next year's sales, which is then used to produce both a point estimate and a 90% confidence interval for the next year's expected sales. The second method is based on a daily bootstrap simulation where the characteristics of each day's sales volume over a historical period is used with a random resampling procedure to produce an expected next year's sales figure. The forecast from this method shows a large gap from the aggregate Monte Carlo method. The convergence of the two methods absolutely depends on the assumptions that are the basis for each. The large gap has to do with what has been assumed about the market. The models themselves determine the numbers based on their assumptions. The final conclusion of this report can be stated as follows: A single estimate of next year's sales figure is not sufficient for an effective annual plan. All possible outcomes must be analyzed and weighed based on the forecasted state of the market. The chain of business data must be monitored closely on a monthly basis and used to adjust resources accordingly.

In addition, the report provides a detailed volatility breakdown by product-category and sales channel, and it identifies that sales volatility is related to transaction structure (the interaction of transaction frequency and order size) and less so with sales share. Accordingly, the report recommends the use of the product-category level and sales channel information for inventory planning and resource allocation.

AdventureWorks two main forecasting models are the two headline models and the two segment-level volatility breakdowns. This report recommends using the Bootstrap model's 90% confidence interval as a baseline for annual budgeting, the Monte Carlo model's confidence interval as a backup for upside scenario, and implement a dynamic monitoring scheme of annual growth rate to appropriately move between the two referencing options. Compared to the traditional point forecasting, this conditional and operationally actionable case framework "provides a more rigorous yet practical basis for the company's risk management."

9. Appendix

- Full code is provided in the File named: Code.ipynb (Jupyter Notebook)
- Random seeds: Monte Carlo simulation seed = 42; Bootstrap simulation seed = 43 (ensures reproducibility)