# 

STAT 112 - Introduction to Data Processing and Visualization Project

The Most Choosen Auto Across Cities and Categories to buy

# by

Bilginalp Büyüktaş

December 8, 2024

# 

# Abstract

# This report provides an in-depth analysis of automobile sales data, highlighting key findings through scatter plots and bar charts. These visualizations were crucial in the study of indicators like overall sales, average prices, and trends for different vehicle categories, geographical regions, and time frames. The data was meticulously cleaned and standardized to ensure accurate and actionable information.

# Introduction

In this research, there was prepared and analyzed a dataset comprising data on 7 varieties of autos with various features. These findings would thus make it easier for the decision-makers to further develop a pricing strategy, build an inventory, and targeting market for business success.

The initial dataset contains various distinct variables:

| *variable name* | *description* | *scale* |
| --- | --- | --- |
| sales | total sale of the auto | interval, continuous |
| priceeach | price of each auto | ratio, continuous |
| date | sale date of the auto | nominal |
| quantityordered | number of units of auto ordered | ratio, discrete |
| dealsize | size of the deal in auto sales | interval, continuous |
| ordernumber | unique number for the auto sale | ratio, discrete |
| status | status of the auto sale | nominal |
| country | country where the auto was sold | nominal |
| phone  productline | auto buyer's contact info  category of the auto sold | ratio, discrete  nominal |

# These categorical columns had irregular string formatting and at least one missing value in each of them. Moreover, the warranty period had two different units of measurement, making it inconsistent. These were rectified in an organized manner during cleaning of the data, which is discussed later in the following sections. Some key insights developed from this exploratory data analysis are elaborated upon in light of certain research questions.

# Automobile sales are pretty dynamic, changing with time trends, customer preferences for choices, and regional variations in sale performance. This report summarizes this very critical metric of total sales, average price, and product line performance using Tableau for effective visualization of data. The dataset contains the category of vehicle, sales region, and patterns over time, hence giving a complete outlook on market dynamics.

# Scatter plots underlined varying relationships of pricing to sales performance that would form the basis for driving actionable outcomes in the optimization of price strategy, inventory management, and marketing activities. Also, regional preference and seasonal trends were depicted using the bar chart. These shall form the basis of driving a data-driven approach toward strategic decisions.

# The following sections discuss in detail cleaning and preparing the data, further followed by the exploration analysis and visualization which have indicated meaningful patterns and trends in the dataset of AutoSales.

# Data Tidying and Cleaning

Data preparation was pursued through consistency cleanup and preparation of the dataset for analysis; cleaning up formatting, standardizing units, outlier detection and handling, duplicate removal, and imputation for missing values. All steps were done with great attention to preserve data integrity and improve the reliability of the analysis.

Handling Missing Values:

The other major activity that was performed in cleaning the data was dealing with the problem of missing values. This is resolved through a focused approach toward the autosales and worlddata2023 datasets through an inner join on the "country" variable of both datasets. In the inner join method, all the records present in the final dataset after cleaning have to match with each other for the values of the country variable.

This was effective for two reasons: it did away with the mismatched entry those entries would have added noise in the analysis if only the consistent data points were allowed to remain, and then integrated more information from worlddata2023 into the autosales dataset and enriched the analytics by providing deeper insights. Last but not least, discrepancy resolution in this stage also spares the subsequent analyses further adjustments due to missing values.

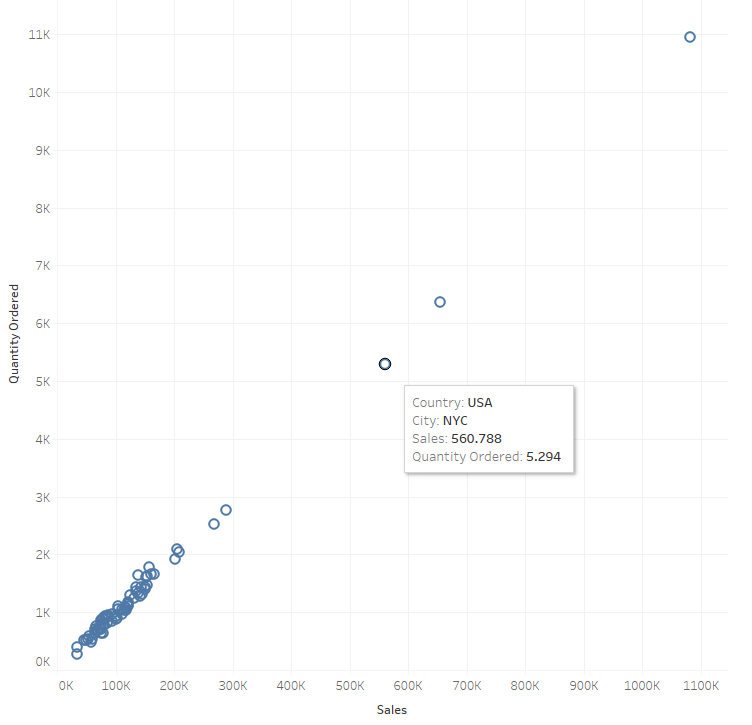
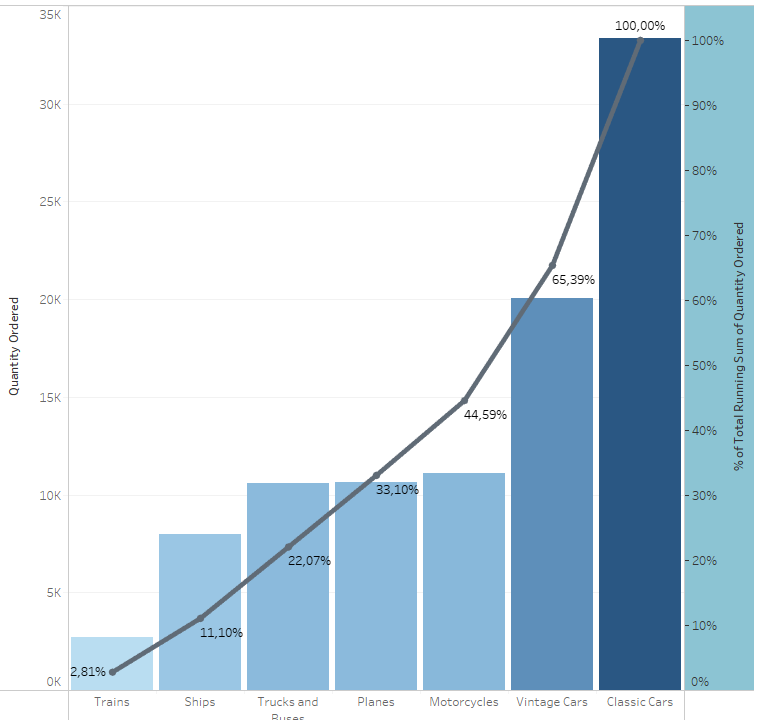
Thus, we could handle the problem of the missing value; the completion of the dataset was done for further analysis.

Duplicate Removal:

It had been cleaned on duplicate entries since that would have led to biased analysis results.  
  
We cleaned and tidied the data into a valid and reliable database, hence making a very strong foundation for analyses and visualizations. The cleaned data set, while analyzed and visualized, yields certain insights and useful conclusions that would be discussed further in the subsequent sections.

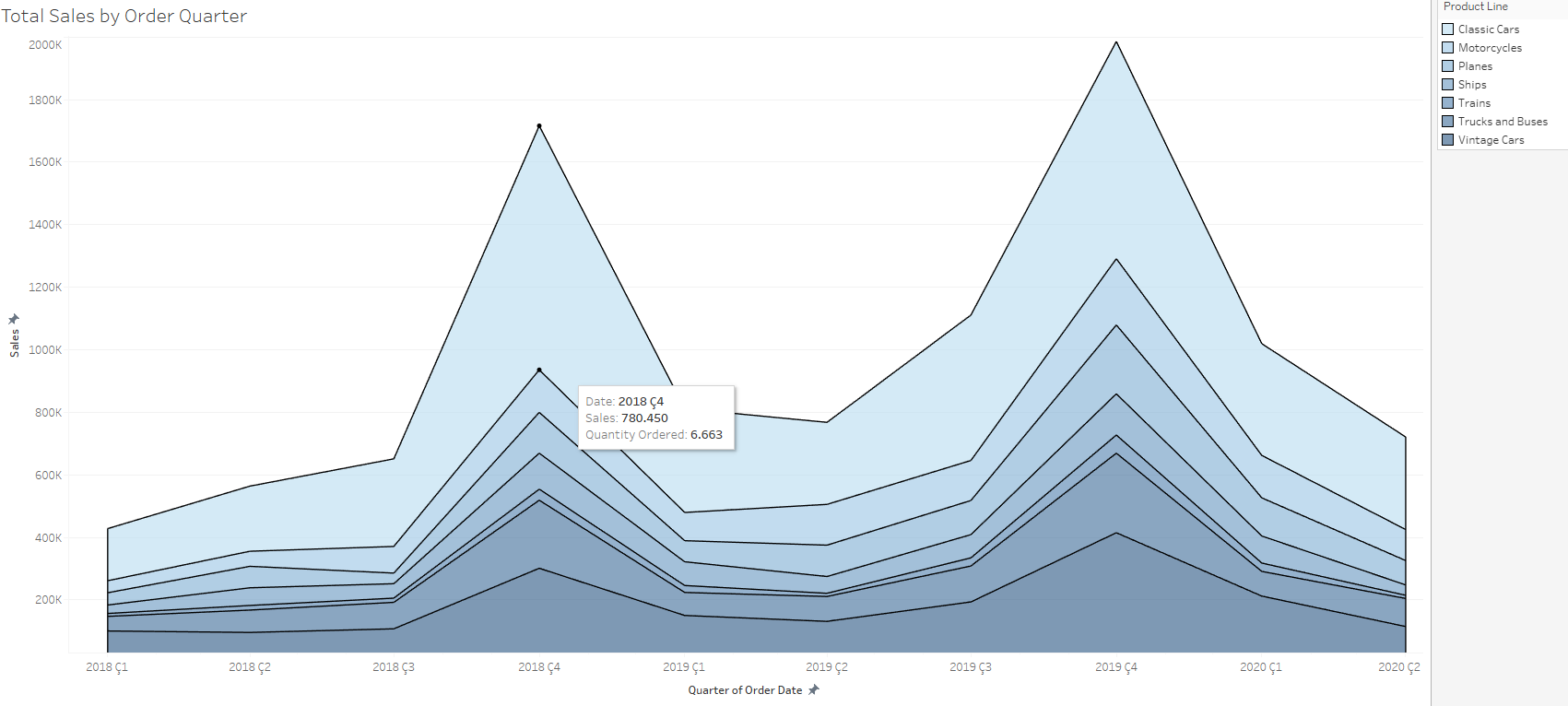
## Research Questions

Which product lines have the highest total quantities ordered, and what factors contribute to their popularity?

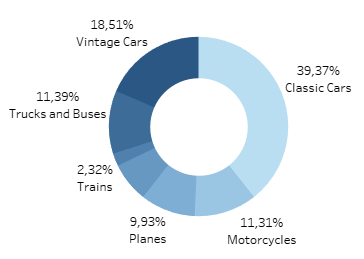
It follows from the chart that the most famous product line is Classic Cars, keeping the cumulative running total of quantities ordered at 100%. The other two product lines in this category are Vintage Cars at 65.39% and Motorcycles at 44.59%, which is a good proportion to attract customers for these products. In contrast, categories like Trains (2.81%) and Ships (11.10%) show rather low quantities ordered, which likely reflects their narrower markets. It is thus probably because of the general appeal of Classic Cars, including the reasons of nostalgia, interest by collectors, and serious marketing efforts, that places them in the leading position, whereas other categories may appeal to some smaller specialized or regional markets.

How do total sales and quantity ordered vary across countries?

Scatter Plot One positive relationship of total sales versus quantity ordered along with some extremes and clustering. This would be for the USA with the City being NYC at 5,294 units ordered for $560,788 in total sales making this a very strong market. Most the points fall below 2,000 units ordered and below $200,000 sales showing that most orders come with small quantities. However, for well-defined anomalies, such as those more than 11,000-unit orders, there is the possibility of studying bulk buying trends or popular products in that region. This distribution does justify targeting high-earning regions like NYC for generating more sales.

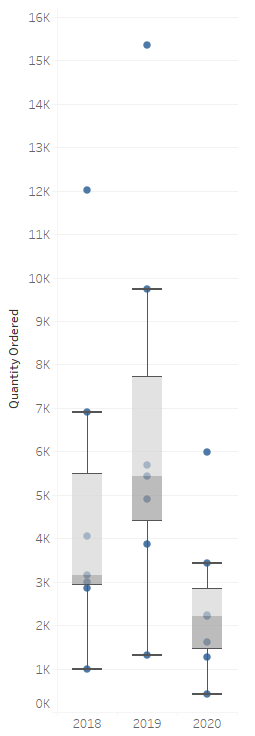
What seasonal patterns can be identified in sales performance across order quarters?

The graph depicts that seasonality strongly affects sales performance, peaking in the fourth quarter of every year, especially in 2018 and 2019. This would most probably be because of high consumer activities during the holiday seasons or at year-end sales promotions, evident in the graph as sales soared to over 1,800K in 2019 Q4. Conversely, sales have always slowed down in the first quarters of the year due to after-holiday slowdowns. Sales of Classic Cars and Vintage Cars have always topped the total in different quarters. It is to be noted that the seasonal pattern depicts Q4 to be quite vital concerning reaping maximum sales; hence, strategic marketing and inventory planning should be done during this period.

Which product line contributes the most to overall sales, and how does its performance compare to other product lines?

That is where the donut chart shows that Classic Cars contribute to 39.37% of sales, hence the most significant product line, followed by Vintage Cars with 18.51% contribution, showing that they strongly stand second in the secondary role. Others were in medium contribution: Trucks and Buses - 11.39%, Motor Cycles 11.31%, and Planes 9.93%. Trains are the least of all with the contribution rate of only 2.32% of it. A big share in itself, Classic Cars would mean wider appeal, and also the consistency of demand, while other contributions will be relatively small to suggest either more niche or region-specific popularity. The given distribution underlines the strategic importance of prioritizing Classic and Vintage Cars when developing sales and marketing efforts.

How has the quantity of products ordered changed over the years, and what factors drive these trends?



The box plot showing yearly variations in the quantity ordered between the years 2018 to 2020. During 2019, the quantity ordered at median value increases, and the range with outliers as high as 16,000 units also characterizes a peak in demand. While in sharp contrast, 2020 saw a deep plunge, with both the median and overall range shrinking tremendously, probably reflecting disruption to the market or reduced consumer activity, possibly due to external factors such as changes in economic conditions or world events. In 2018, the demand is more average, with less volatility compared to 2019. This, therefore, means that an excellent stretching of the market occurred in 2019 and a contraction in the year 2020; this would mean further investigation into the commodity's availability, marketing strategies, and the prevailing economic circumstances that could have brought about this change.

# Conclusion

This analysis gives an overall view into the autosales dataset; visualizations of higher order have been made to allow for the most important determination of patterns and trends in the data. From the findings, Classic Cars and Vintage Cars emerge as the two best items in sales, with their respective performances based on regional and seasonal factors. This could also be complemented by seasonality in the sales peak in the fourth quarter of each year, especially for the year 2019 and the big decline in 2020 to show some external market challenges.

A positive correlation between pricing strategies and sales performance, supplemented by regional preferences in terms of product line popularity, would greatly enhance targeting marketing efforts and strategic inventory management. In addition, the yearly quantity ordered supports the fact that one must move with the dynamics of the market, which changes every year, hence building resilience in fluctuating demand.

If the region that pulls through best, product availability, and other such inputs are optimized, it will help in fine-tuning the business for profitability while matching up to customers' emerging demands.