Data Science EDA Test Report

Table of Contents

Introduction

The dataset provided for this test contains data related to the sale of machines, particularly construction and heavy equipment. The sheet contained missing values, outliers, and inconsistencies in data type; typical challenges faced by data scientists. Our initial objective was to prepare the data for analysis, followed by an in-depth exploration of the relationships between various variables. Subsequently, I built a machine learning model to predict a target variable, 'YearSold,' and evaluated its performance using the Mean Absolute Error(mae), Root Mean Squared Error(rmse) and R-squared(r2) matrics.

Task 1: Data Cleaning

I performed data cleaning on the provided dataset, which contained missing values, outliers, and inconsistencies. Here are the key steps I took:

1. Loaded the dataset and performed exploratory data analysis (EDA) to understand the structure of the data.

    - The original Dataset had a structure of the Number of rows and columns: (412698, 54).

2. I Identified and handled missing values appropriately.

    - After dropping columns with mixed data types that were not relevant to the analysis, the final data set had a structure of the Number of rows and columns: (373307, 16).

3. Detected and handled outliers by considering domain knowledge and statistical methods.

- By using the scipy package to eliminate outliers that fell within the outer boundaries of -3 and 3 standard deviation.

4. Ensured data types were appropriate for analysis by converting date columns and encoding categorical data

    - I extracted the year from the date in the “saledate” column and formed “SaleYear”, which comprised of only sale years.

Task 2: Data Analysis

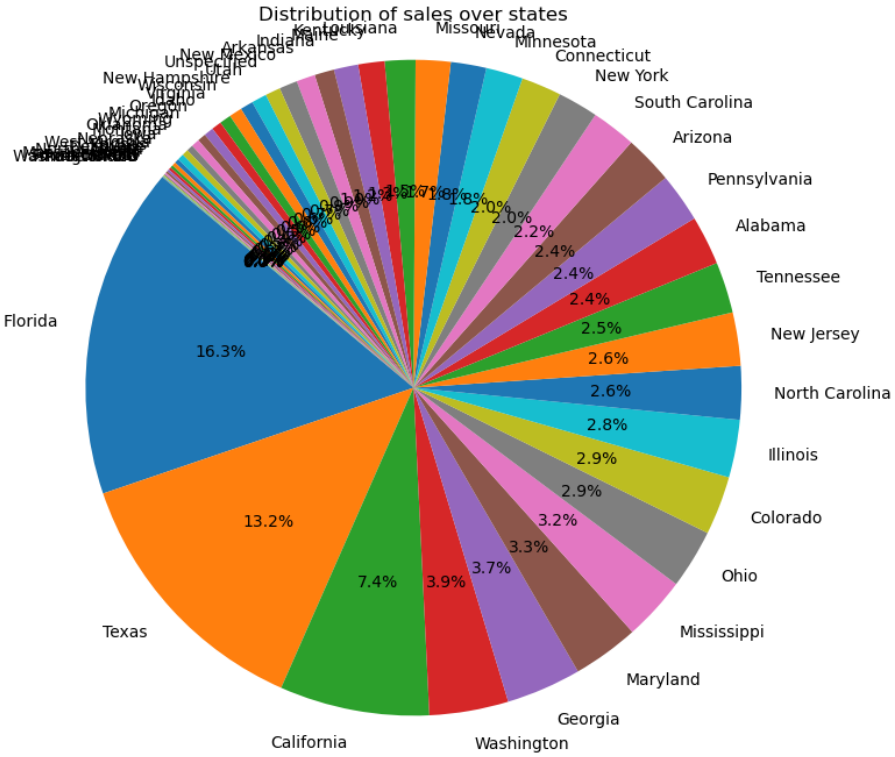
In Task 2, I focused on data analysis using the cleaned dataset. Here's what was done:

6. Conducted univariate analysis on the ‘state’ variable to understand its distribution and characteristics.

    - I Identified unique values for each variable, to better select a variable for analysis.

7. Visualizations And Insights

    - Pie plot of Distribution of sales over states

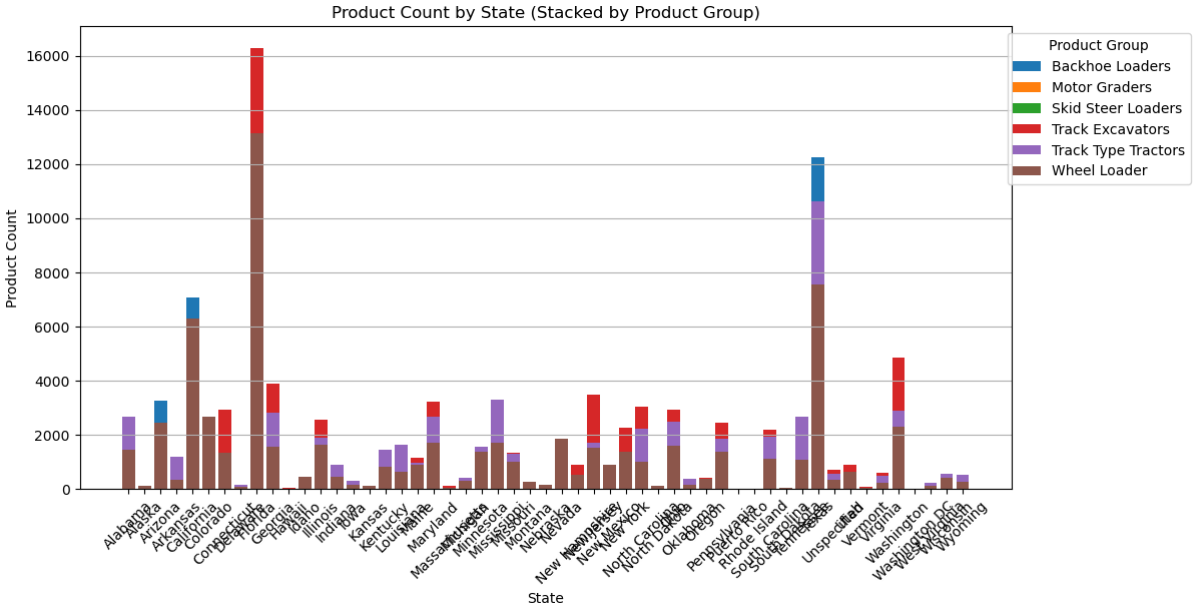


Insight 1:

a. For the given period, Florida made the highest number of sales contributing to 16.3% of the total number of sales for the period.
b. Texas sold the second largest number of heavy machines, equating to 13.2% of all sales.

c. The top 8 states made more sales than the rest of all the other states combined.

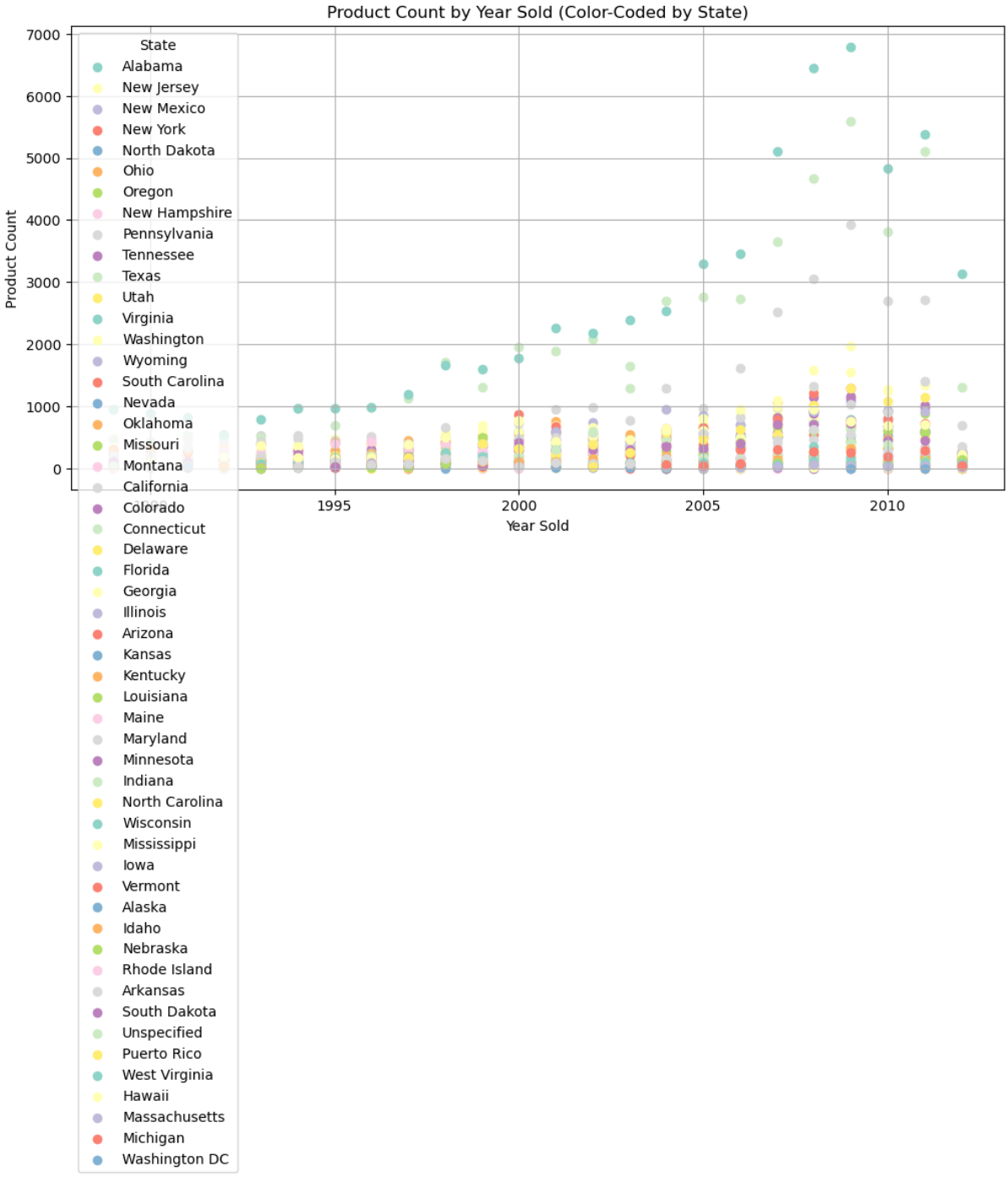
- Created a Stacked Bar chat of product count by state, with each bar showing its proportionate product constitution in its overall sales by state.



Insight 2:

a. The ‘wheel loader’ was the most sold product in all states, constituting more than 50% of sales in each state.
b. The second most sold product was the ‘Track Type Tractors’. Followed by the ‘Track Excavarotors’.

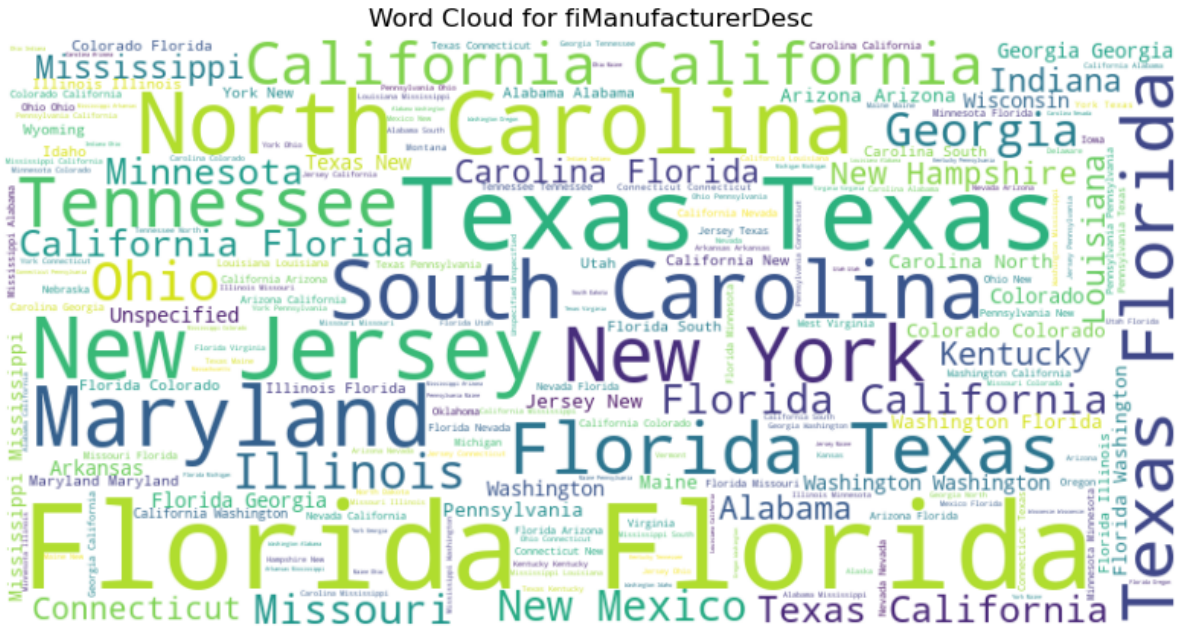
- Scatter Plot of product count by Year sold color coded by the states the products were sold in.



Insight 3:

a. More Products were consistently sold in Florida over the given period.
b. Most products across all states were sold between 2006 and 2012.
c. Sales prior to 1995 were limited across board.

- Word cloud for states

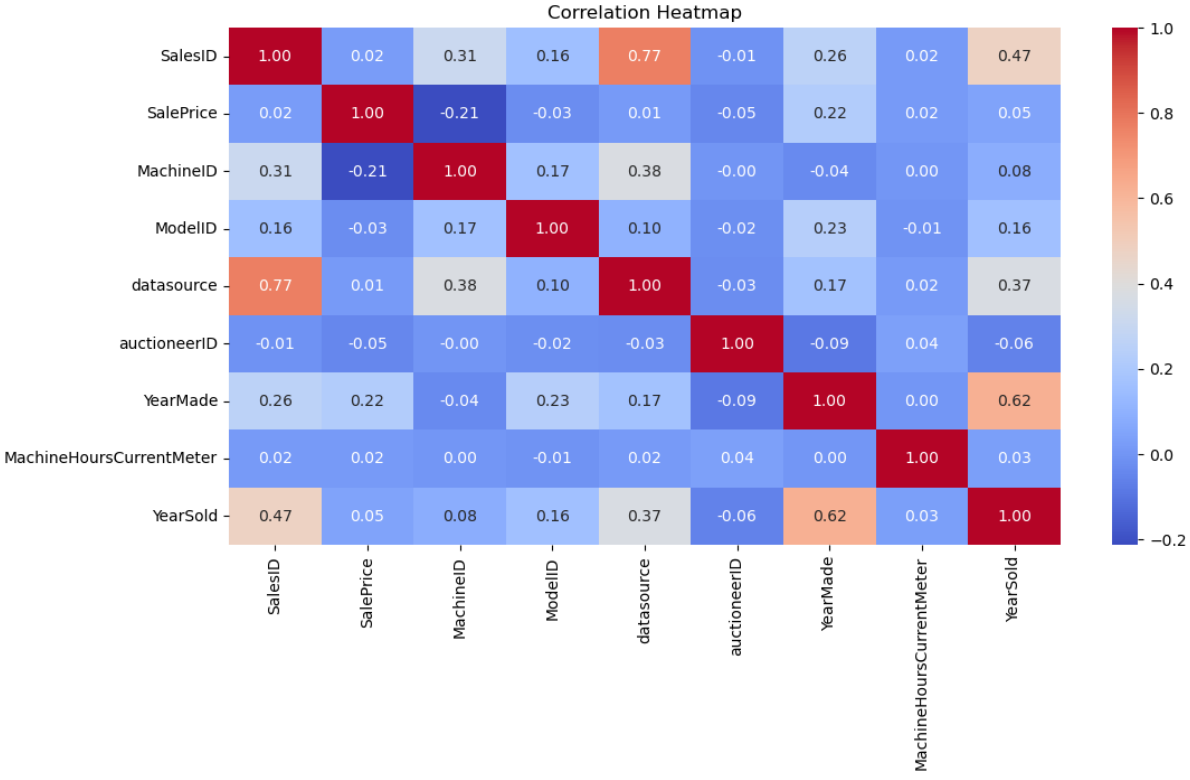


Insights 4:

a. Visual representation of sales of heavy machinery prominence in all states

8. Performed correlation analysis to identify relationships between variables,

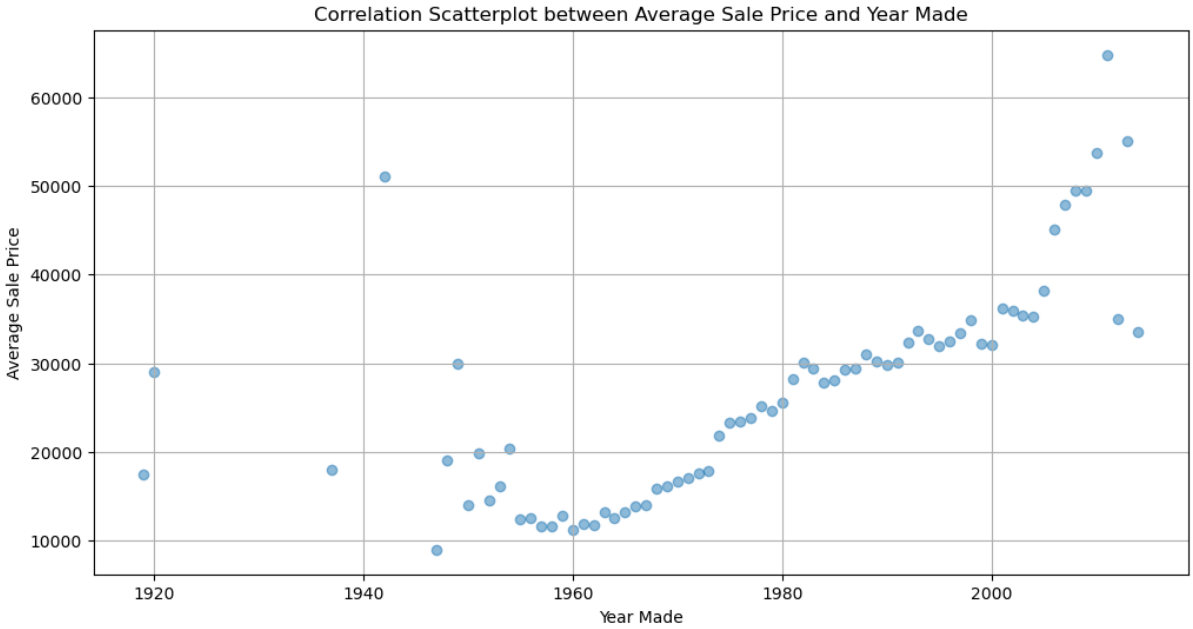
- Heat Map correlation to identify variables with strong correlation.



Insight 5:

a. The Strongest correlation between the quantitative part of the dataset lies between 'SalesID' and 'Datasource' variables with an index of 0.77.
b. Second strongest correlation lies between 'year Made' and 'year sold', Hence Our Model will attempt to predict the year a machinery will be sold given other variables with an index of 0.62.
c. Shows all correlation between all quantitative variables.

- Scatter plot of Average Price with year made.



Insight 6:

a. There is some correlation between average price and year Made.

- Scatter plot of Average Price with year sold.

Insight 7:

a. There is very little correlation between average price and year sold.

Task 3: Machine Learning

For Task 3, I built a basic machine-learning model to predict the year sold target variable. This can predict when machinery is going to be sold with a very good degree of accuracy.
The test size was set to 20% and the training size was 80%.

9. Choose a suitable machine learning algorithm and explain why you selected it.

- The machine learning model I used was the **Random Forest Regressor model.**

- **I used it because** it can capture non-linear relationships between the features and the target variable which is essential because real-world data such as the one provided is not linear. It also combines multiple decision trees to make predictions, which helps to reduce overfitting and improve predictive accuracy. Since I aimed to predict the 'YearSold' (a numerical value), a regression task was required, and Random Forest is well-suited for regression problems.

- Random Forest provides feature importance scores, which can help you identify the most influential features in predicting the target variable ('YearSold').

10. Evaluated the model's performance on the testing data using appropriate metrics, including Mean Absolute Error, Root Mean Squared Error, and R-squared (R2).

Results:

```
Mean Absolute Error: 0.5920604638810167
Root Mean Squared Error: 0.9492846452153553
R-squared (R2): 0.8802925213838653
```

Insights Gained:

- **Mean Absolute Error (MAE):** This measures how close the model's predictions are to the actual values. In this case, the model's predictions are, on average, about **0.592** years away from the real 'YearSold.' So, if the model predicts a machine was sold in 2000, the actual sale year is typically around 0.592 years different, which is quite accurate.

- **Root Mean Squared Error (RMSE):** RMSE is similar to MAE but gives more importance to larger errors. Here, it's about **0**.**949**. This means that while most predictions are very close to the real values, there might be a few predictions with larger differences. It's a good sign that the model's errors are consistent.

- **R-squared (R2):** R2 measures how well the model fits the data. In this case, the model explains about **88%** of the differences in 'YearSold.' This is great because it means the model is doing an excellent job of understanding and predicting the year a machine was sold.