BAN 620 Case Assignment 1

TEAM 1

SHRIYA ARORA – sz9461

PALLAVI NAIR- fb4097

PRIYA MADHUSUDANAN- hy1162

ADITYA MANE- yy5910

SAI KUMAR- lp2345

HARSHITHA ANANTHULA- pi3407

CAL STATE EAST BAY

Case 1: Carvana Case

Introduction: The used car market is a profitable segment of the automobile industry. Carvana, a pioneer in the online used car marketplace, increased its revenue to $13.6 billion driven largely by an increase in used car sales. Purchasing a used car can be attractive to customers who are looking to save money. However, buying a used car can be a complex process, with many factors to consider such as cost range, pre-purchase inspection, ownership validation, vehicle history report, age, and mileage.

We aim to address this problem by analyzing various factors behind a car purchase to determine whether a particular auction purchase is a good or bad buy. Our goal is to analyze the 73,000 transactions from Carvana and develop a prediction model to help us determine whether a specific transaction is a good or bad buy.

Data Source: The dataset was obtained from Kaggle. The dataset was originally provided by Carvana, a Technology business start-up in Tempe, Arizona.

Goal: The key here is to analyze the 73K transactions from Carvana and come up with the prediction model that helps us understand whether a particular transaction is a Good/ Bad buy.

Dataset Analysis:

a) The dataset initially had 34 attributes with the target attribute being - IsBadBuy
b) The dataset had 19 numerical attributes and 15 categorical attributes.

```
In [6]: car_df.shape
Out[6]: (72983, 34)
```

c) Attributes PRIMEUNIT and ACGUART had only less than 1% percent of data and did not help enough with model building and were removed from the analysis. The visual representationcan be seen in Fig.1
We also removed the numerical variables VNZIP1 and RefId.
VNZIP1 was the zip codes of where the cars were sold. This is redundant because, we can infer this from the variable VNST (state codes). RefId is just a transaction ID which had no meaning.

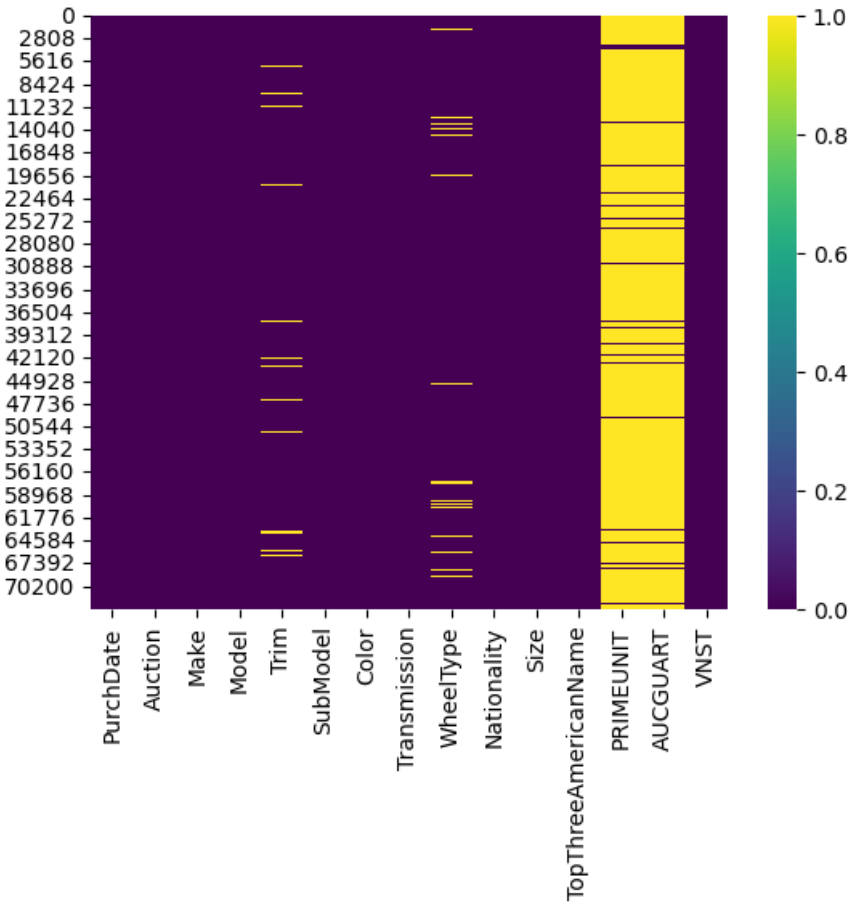


Figure 1:Heat Map showing missing data values.

d) As per the initial analysis shown below, we can see the overall % of bad buy:

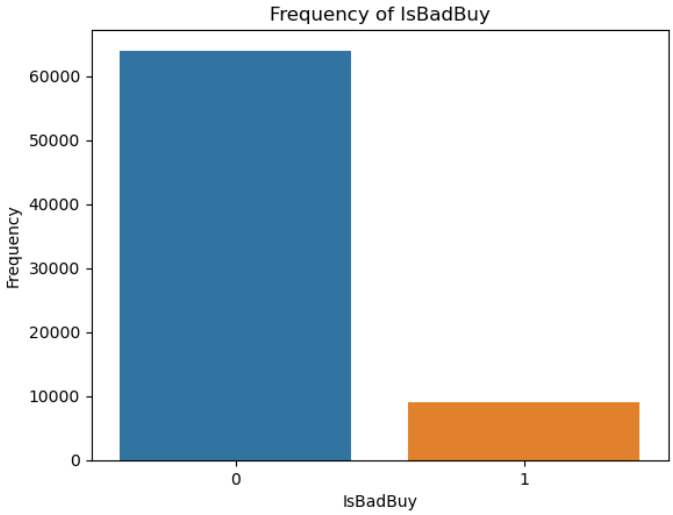
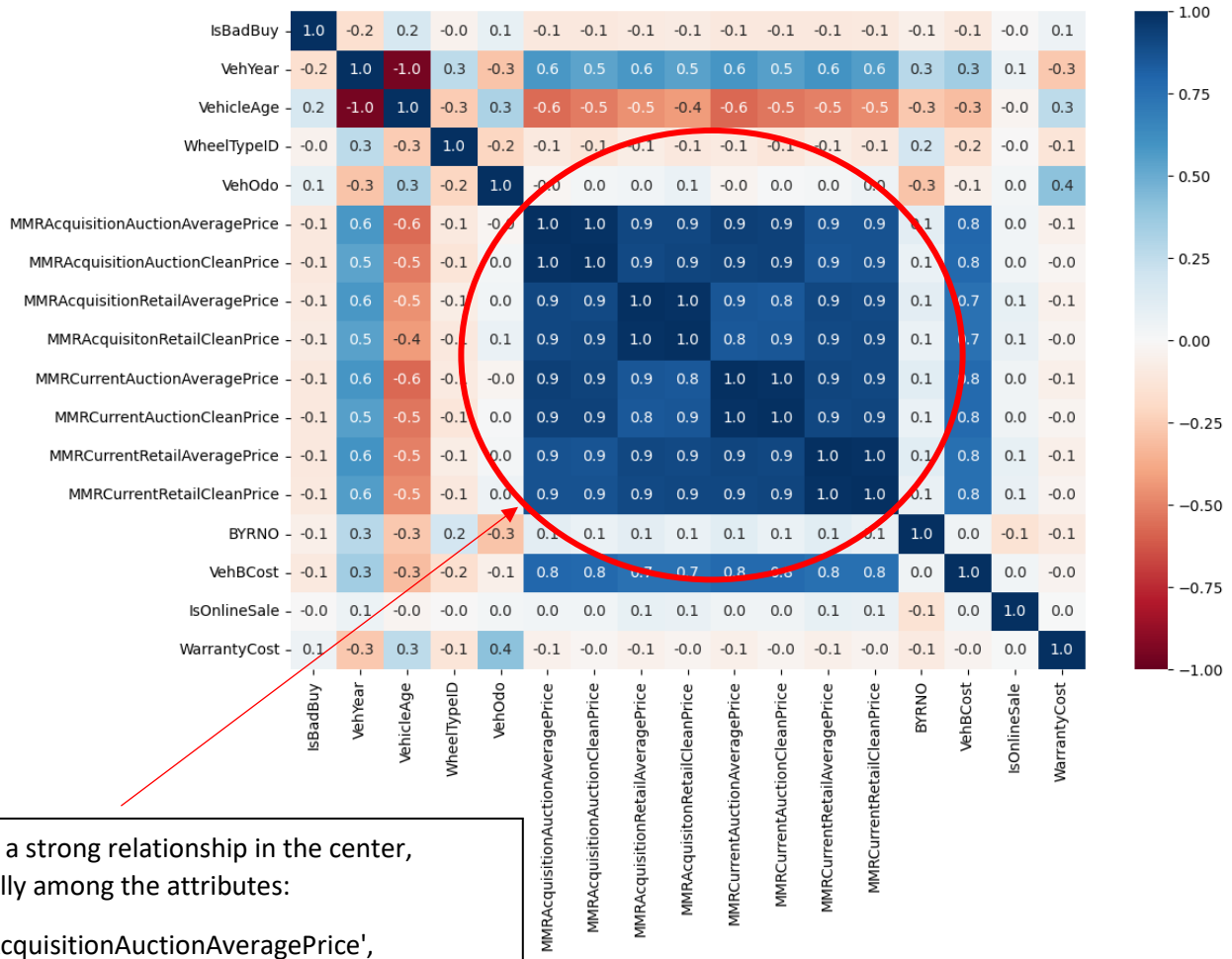


Figure 2: Not a Bad Buy: **87.7%,** Is Bad Buy: **12.3%**

e) Correlation Matrix Heat Map:



We see a strong relationship in the center, especially among the attributes:

'MMRAcquisitionAuctionAveragePrice',
'MMRAcquisitionAuctionCleanPrice',
MMRAcquisitionRetailAveragePrice',
'MMRAcquisitonRetailCleanPrice',
MMRCurrentAuctionAveragePrice',
'MMRCurrentAuctionCleanPrice',
MMRCurrentRetailAveragePrice',
'MMRCurrentRetailCleanPrice'

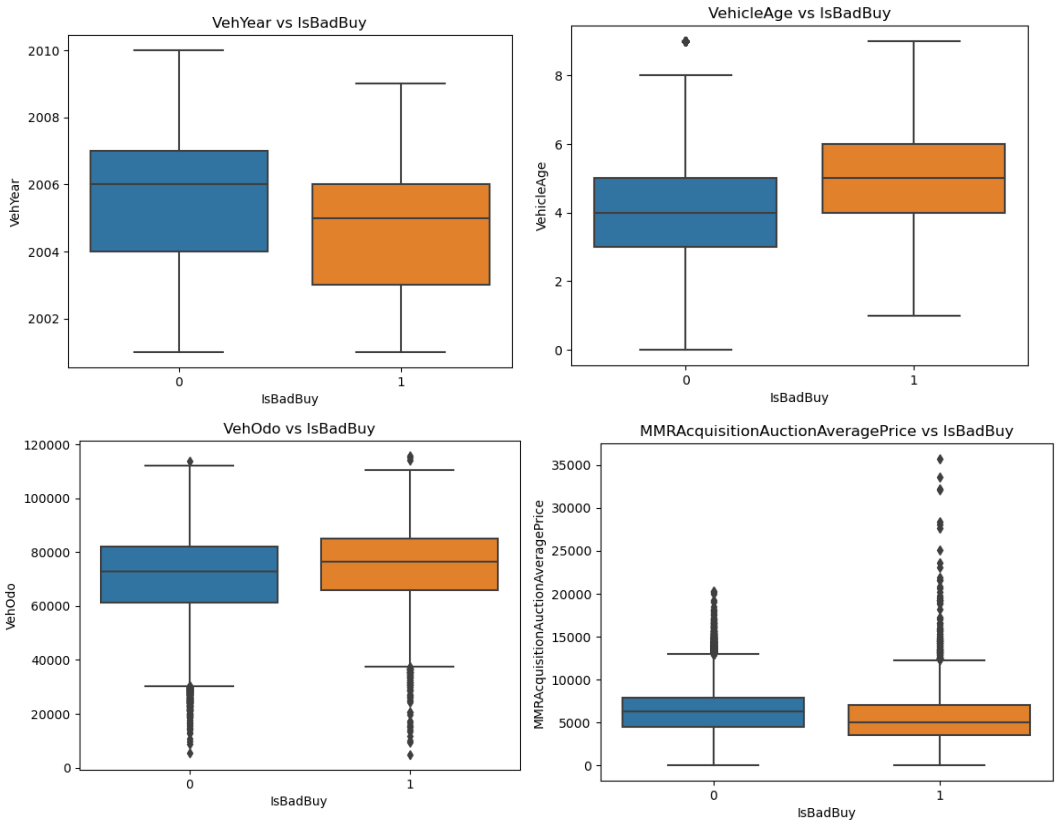
Analysis on some Leading Indicators of a Bad Buy: Exploratory Data Analysis

a) In our analysis, we plotted box plots to initially identify if there is any relationship between IsBadBuy and other variables.

Based on the box plots we were able to identify the below variables:

1. Vehicle Age: The years elapsed since the manufacturer's year.
2. Vehicle Year: The manufacturer's year.
3. MMR Acquisition Auction Average Price: Acquisition price of the vehicle in average condition at the time of purchase.
4. Vehicle Odometer: The vehicles odometer reading.

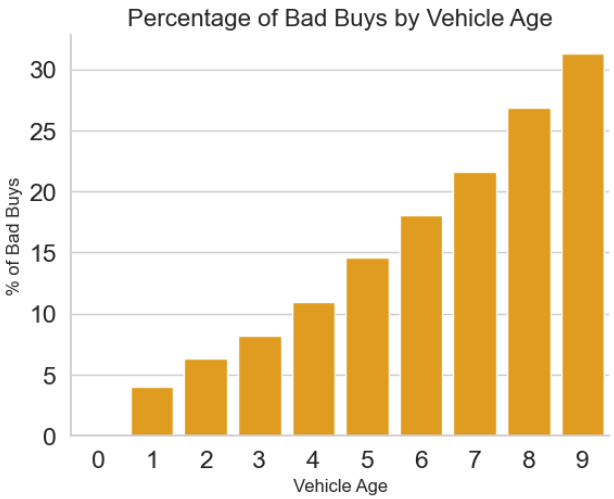
Shown below are the box plots, where by the distribution of the data for IsBadBuy(0 & 1); we can see that the above mentioned variables shows dependency on IsBadBuy:



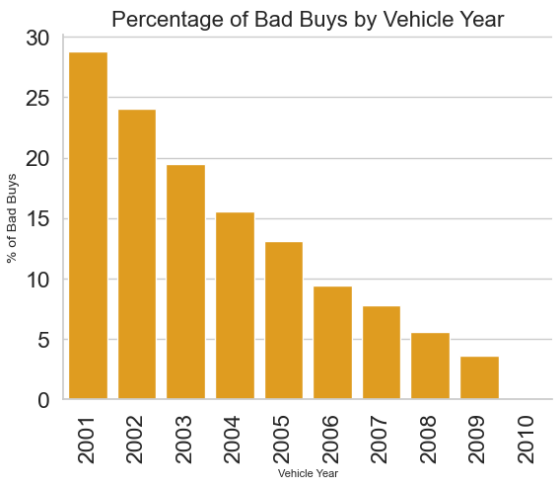
By visually comparing the boxes and whiskers for different categories, we can determine the differences in the distribution of the data.

Further, we went on further analyzing the relationship and effectively identify the indicators. We plotted bar plots of IsBadBuy v/s the identified variables:

VEHICLE AGE



VEHICLE YEAR

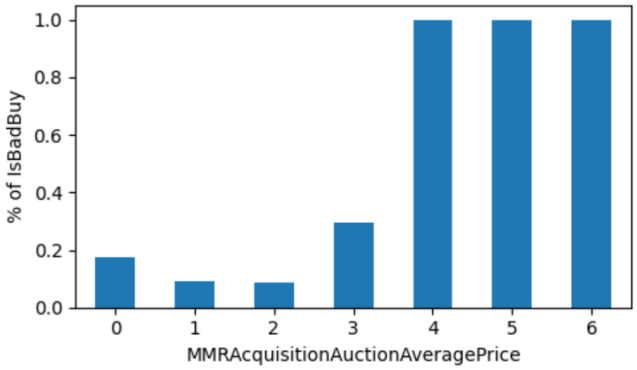


"As we can observe from the above chart, as the Vehicle Age increases, the probability of a Bad buy also increases. In addition to that, since Vehicle Age and Vehicle Year are highly negatively correlated (-0.96), we can observe a relationship between Vehicle Year and Is Bad Buy as well".

ACQUISITION AVERAGE PRICE

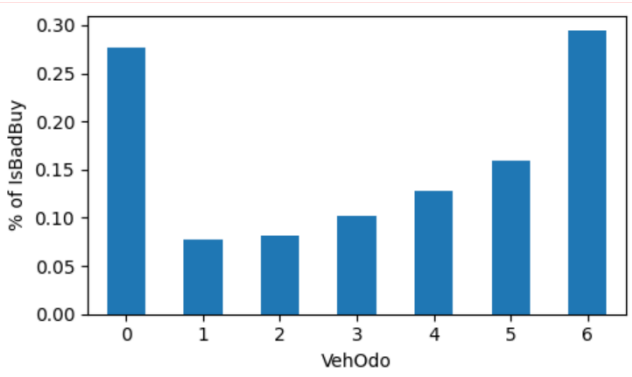
VEHICLE ODODMETER

As the MMR Acquisition Auction Average Price Increases, the Percentage of Bad Buy increases.



Categories-MMRAcquisitionAuctionAveragePrice: [(-35.722, 5103.143] < (5103.143, 10206.286] < (10206.286, 15309.429] < (15309.429, 20412.571] < (20412.571, 25515.714] < (25515.714, 30618.857] < (30618.857, 35722.0]]

Note: For visual representation we have created categories for MRAcquisitionAuctionAveragePrice & Vehicle Odometer



Vehicle Odometer exhibits a relationship like Vehicle Age vs Bad Buy. The only difference we can infer is that even a few cars which have an Odometer reading in the range of 4714.108, 20666.714 have about 27% vehicles marked as Bad Buy.

Categories- Vehicle Odometer: [(4714.108, 20666.714] < (20666.714, 36508.429] < (36508.429, 52350.143] < (52350.143, 68191.857] < (68191.857, 84033.571] < (84033.571, 99875.286] < (99875.286, 115717.0]]

Based on our visual analysis, our opinion is that:

1. Vehicle Age: The years elapsed since the manufacturer's year.
2. Vehicle Year: The manufacturer's year.
3. MMR Acquisition Auction Average Price: Acquisition price of the vehicle in average condition at the time of purchase.
4. Vehicle Odometer: The vehicles odometer reading.

Are the major indicators of a bad buy.

Overall Recommendation:

Based on the analysis we have done, some recommendations for Carvana to make fewer bad buys are:

1. Focus on acquiring newer cars with lower vehicle age.
2. Be cautious when purchasing vehicles with high mileage and thoroughly inspect them to ensure they are in good condition before making a purchase. Consider setting a maximum mileage threshold for vehicles they purchase to reduce the likelihood of buying a bad buy.
3. Be mindful of make and model: certain makes and models may be more prone to issues and have a higher likelihood of being a bad buy.
4. Consider market demand: if a car is in high demand, there may be a higher likelihood of buyers overlooking potential issues and therefore a higher likelihood of bad buys.
5. Be wary of pricing: if a car is priced too high or too low, it may be an indicator of potential issues or hidden problems.

Conclusion:

By analyzing various factors behind a car purchase, we were able to identify the most significant indicators of a bad buy. Carvana can use these indicators to make better-informed decisions when purchasing vehicles and reduce the likelihood of buying a bad buy.

Case 2: Airfare Case

Introduction: The Airfare problem occurred in the late 1990s in the United States, where major cities were facing issues with airport congestion due to the 1978 deregulation of airlines. This deregulation freed both fares and routes from regulation, allowing low-fare carriers like Southwest (SW) to compete on existing routes and begin nonstop service on routes that previously lacked it. While building new airports was not a feasible solution, sometimes decommissioned military bases or smaller municipal airports could be reconfigured for commercial use.

Many stakeholders, including airlines, city, state and federal authorities, civic groups, the military, and airport operators, were involved in the issue. An aviation consulting firm aimed to provide advisory contracts to these players, and predictive models were needed to support its consulting service. One crucial prediction the firm aimed to make was the fares, in case a new airport was brought into service. To accomplish this, the firm utilized Airfares.csv, a file containing real data collected between Q3-1996 and Q2-1997. *(The Airfare Problem: Predicting Fares for a New Airport., 1999) (Shmueli, 2017)*

Data Source: The dataset was obtained from Kaggle.

Goal: The primary focus is to analyze the 638 transactions from the Airfare data set and come up with the prediction model that helps us understand what are the major influencing factors for an Air ticket price, and to devise a model with best possible accuracy to predict the Air fare.

Dataset Analysis:

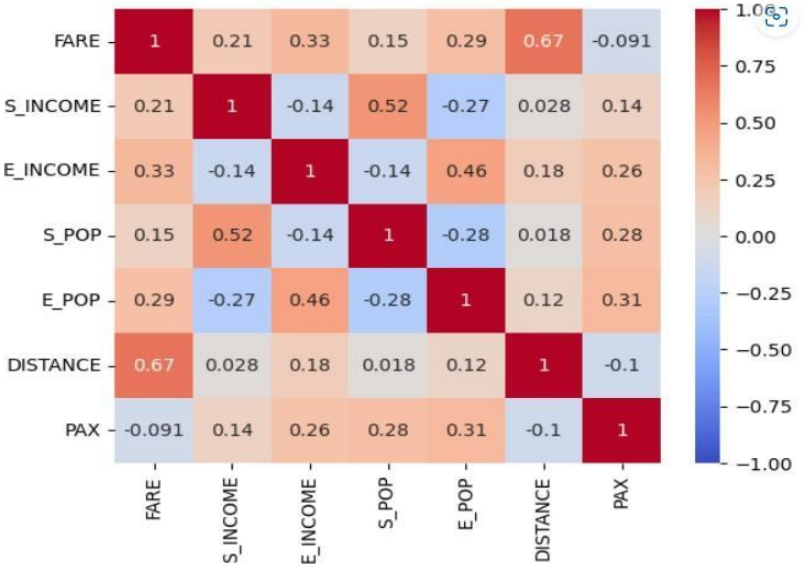
- The dataset initially had 18 attributes with the target attribute being – Fare. In our preliminary data cleaning, we had dropped the first four categorical variables.
- The dataset had 10 numerical attributes and 4 categorical attributes.

```
airfares_df.shape
```

```
(638, 14)
```

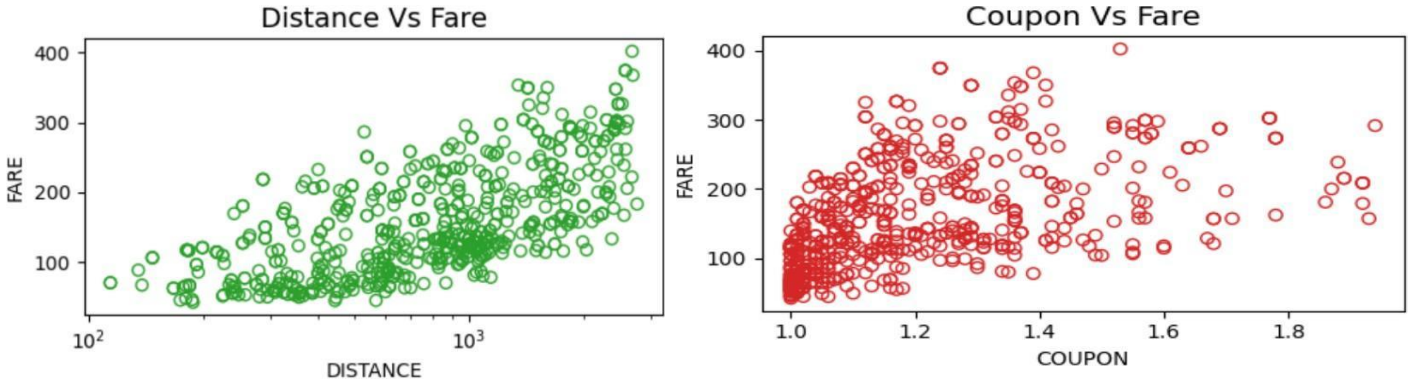
a) Relationship between Airfares and other numerical predictors

To show relationship and correlation of Airfares with other numerical variables, we represented that using a heatmap and a correlation table.



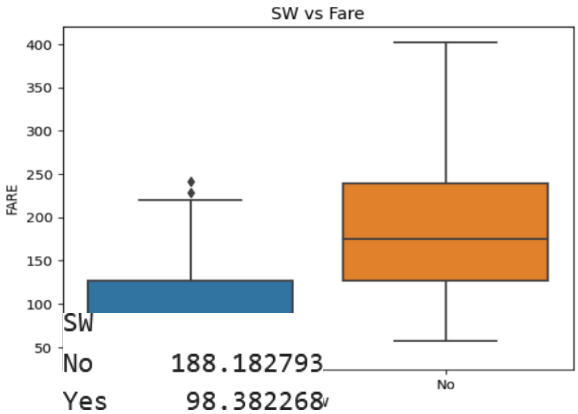
FARE	1.000000
DISTANCE	0.670016
COUPON	0.496537
E_INCOME	0.326092
E_POP	0.285043
S_INCOME	0.209135
S_POP	0.145097
NEW	0.091730
HI	0.025195
PAX	-0.090705

As we can observe from the Heatmap and the correlation table, the numerical variable FARE has a higher correlation with **Distance**, followed by weak relationship with **Coupon and E_Income**.



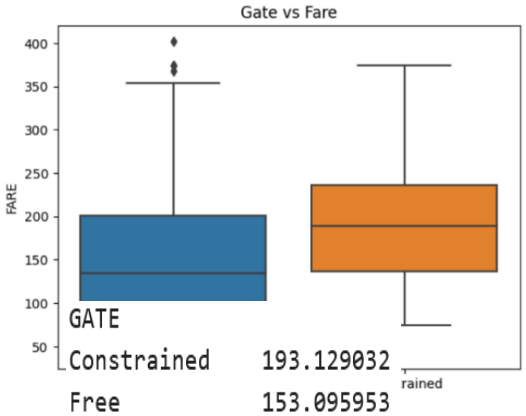
The relationship between Fare vs Distance and Coupon and represented using scatter plots. The spread of the datapoints visually can be interpreted. As we can see, a linear relationship between Fare and Distance; and a weak linear relationship between Fare and Coupon is visible.

b) Effect of categorical variables on Airfare



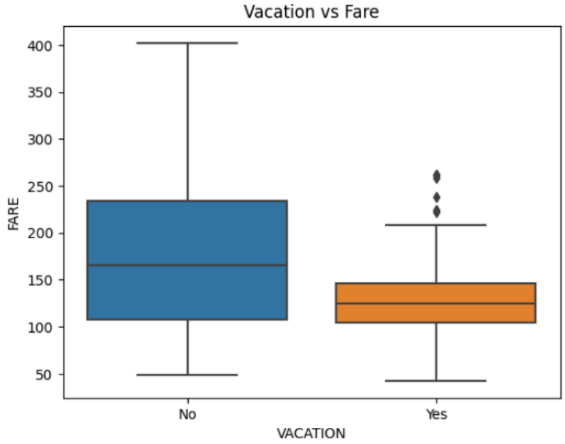
SW

No	188.182793
Yes	98.382268



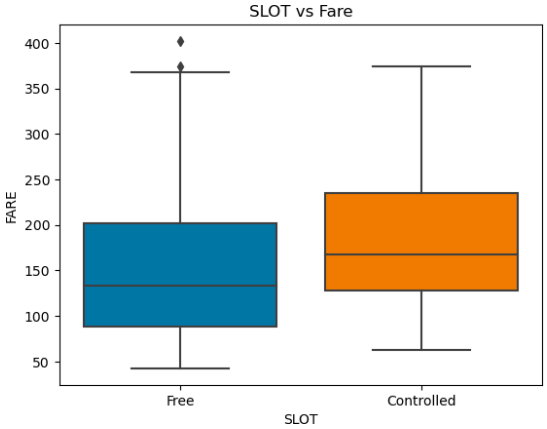
GATE

Constrained	193.129032
Free	153.095953



VACATION

No	173.552500
Yes	125.980882



SLOT

Controlled	186.059396
Free	150.825680

Box plot interpretation:

- From the above Box plots, we can infer that the average Airfare at the Gate which is constrained is about 25% more than the average Airfare at the Gate which is Free.
- The average Airfare with SW as No (South West Airline does not serve the route) is 52% more than the average Airfare with SW as Yes (South West Airline serves the route).
- Vacation Tickets are priced 40% less than non- Vacation Tickets.
- Average Airfare at the Slot which is Constrained is about 23% more than the average Airfare at the Slot which is Free

c) Prediction Model for Airfares

Data set: Primarily, we divided the data set into training set: 60% and validation set: 40%

Predictors: We have used a model consisting of 6 variables as predictors and Fare as the dependent variable.

The Predictors are Distance, E_Income, Gate, Slot, SW and Vacation. *(Independent Variables)*

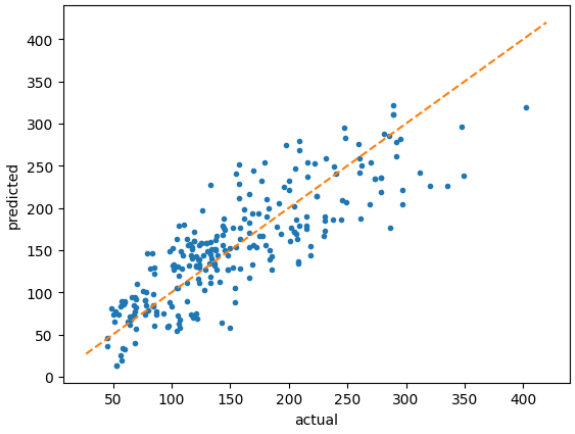
```
intercept  115.58334471028749
      Predictor  coefficient
0      DISTANCE     0.075886
1      E_INCOME     0.001296
2     GATE_Free   -27.915692
3     SLOT_Free   -16.995218
4        SW_Yes   -50.557866
5  VACATION_Yes   -55.514109
```

```
Regression statistics

                    Mean Error (ME) : 0.0000
         Root Mean Squared Error (RMSE) : 38.9298
              Mean Absolute Error (MAE) : 30.4902
            Mean Percentage Error (MPE) : -5.3841
Mean Absolute Percentage Error (MAPE) : 22.6308
```

The above statistics provide the intercept and the co-efficient of the independent variables which constitute the model. Statistics provides the error metrics of the model on the validation data set.

A scatterplot is plotted against actual vs predicted values, the trendline is observed.



OLS Regression Results

Dep. Variable:	FARE	R-squared:	0.755
Model:	OLS	Adj. R-squared:	0.751
Method:	Least Squares	F-statistic:	192.7
Date:	Thu, 06 Apr 2023	Prob (F-statistic):	2.80e-111
Time:	09:07:29	Log-Likelihood:	-1940.8
No. Observations:	382	AIC:	3896.
Df Residuals:	375	BIC:	3923.
Df Model:	6		
Covariance Type:	nonrobust		

d) Yes, the developed model to predict FARE can be helpful for airlines when a new airport is brought into service.

1. The model can provide valuable insights into how different factors such as distance, coupon, vacation, income, population, and other numerical and categorical predictors affect the fare prices.
2. Airlines can use this information to make informed decisions about pricing their Tickets for the new airport.
3. By analyzing the relationship between fare and other numerical predictors, the model can help airlines understand which factors have a significant impact on fare prices. For example, airlines can determine how changes in distance or income levels may affect fare prices and adjust their pricing strategy accordingly.
4. Additionally, by analyzing the effect of categorical predictors on fare using various tables, airlines can gain insights into how factors such as city, airport type, and slot availability influence fare prices.
5. The developed model can also be used to predict fare prices for the new airport based on the available data. By using the accuracy measures based on validation data, airlines can assess the performance of the model and make adjustments as needed. This can help them estimate appropriate fare prices for the new airport, taking into consideration the specific characteristics of the airport and the surrounding area.

However, it's important to note that the model's accuracy and usefulness may depend on the quality and relevance of the data used for training and validation.

1. If the new airport and its surrounding area have significantly different characteristics compared to the data used for model development, the model may not accurately predict fare prices. Therefore, airlines should exercise caution and validate the model's predictions with real-time data from the new airport before making final pricing decisions.
2. The model should be treated as a tool for guidance rather than a definitive source of fare pricing for the new airport.
3. Additionally, other factors such as competition, market demand, and external economic factors should also be taken into consideration when setting fare prices for a new airport.

Conclusion: Overall, the model can provide valuable insights for airlines in setting fare prices for the new airport, but it should be used in conjunction with other factors and real-time data for accurate pricing decisions. It is recommended to use a combination of data-driven models and expert judgment for pricing decisions in the airline industry.

References:

The Airfare Problem: Predicting Fares for a New Airport. **Smith, J. D. 1999.** 1999, Journal of Aviation Management and Education, 1(1), , pp. 25-36.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. 2017. *Data Mining for Business Analytics: Concepts, Techniques, and Applications.* s.l. : R. Wiley., 2017.