BAN 620 Data Mining

Assignment 2

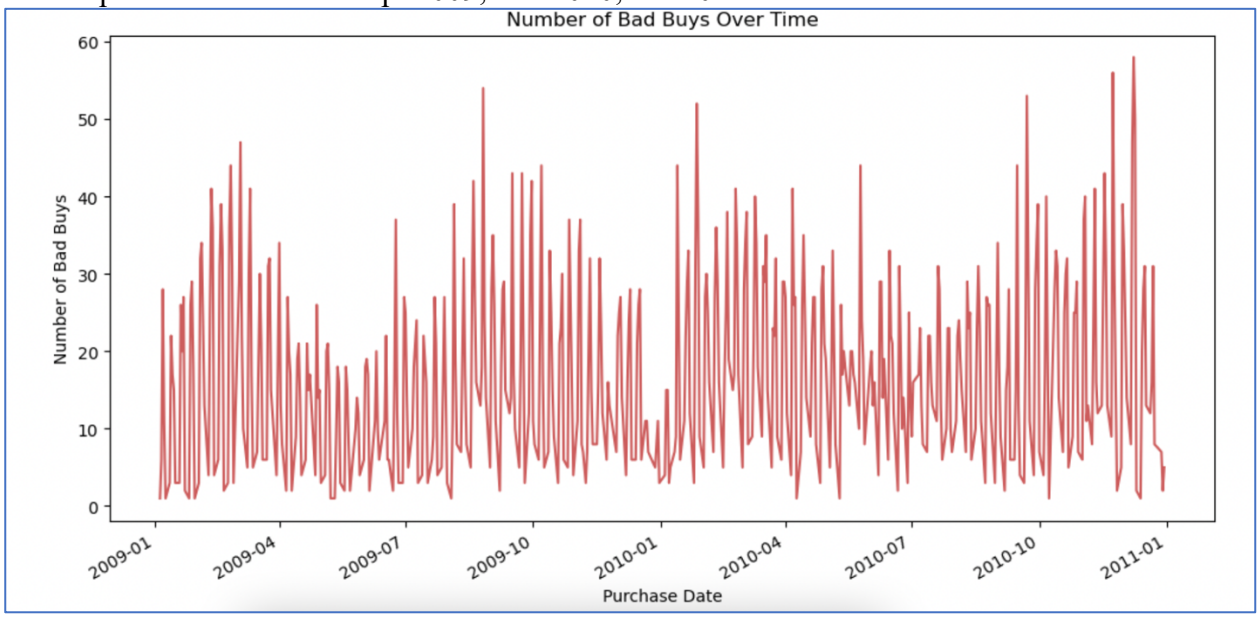
Submitted by: Group 4

Name	NETID
Jasvitha Buggana	js3225
Krupa Shah	yc4954
Manan Upadhyay	rs6739
Preksha Shah	cz2412
Shivani Agrawal	lw3758

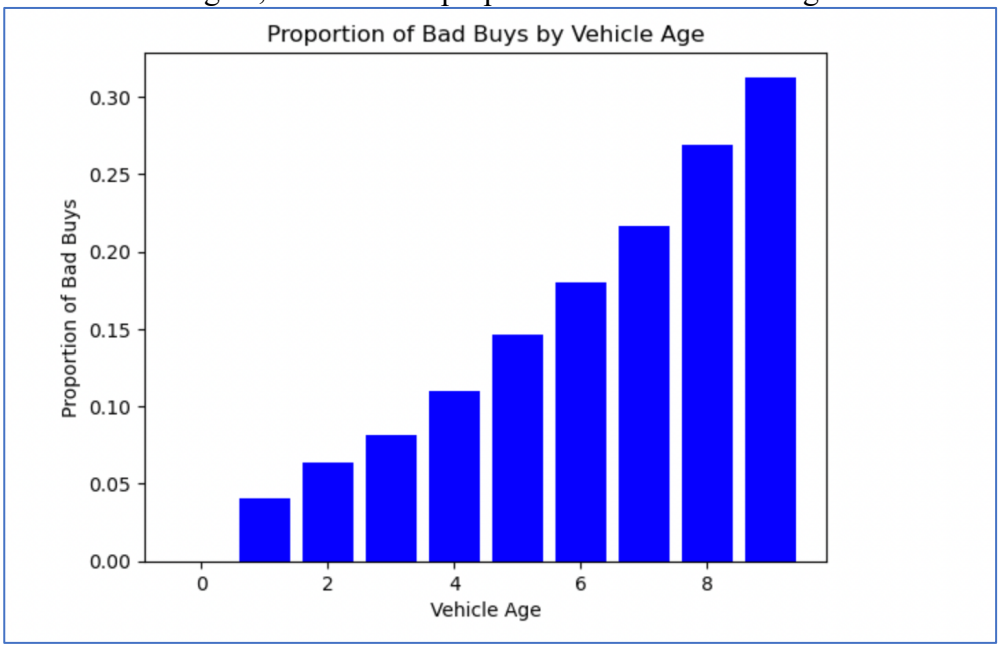
Carvana case:

a. Using visual analytics to identify some leading indicators of a bad buy.

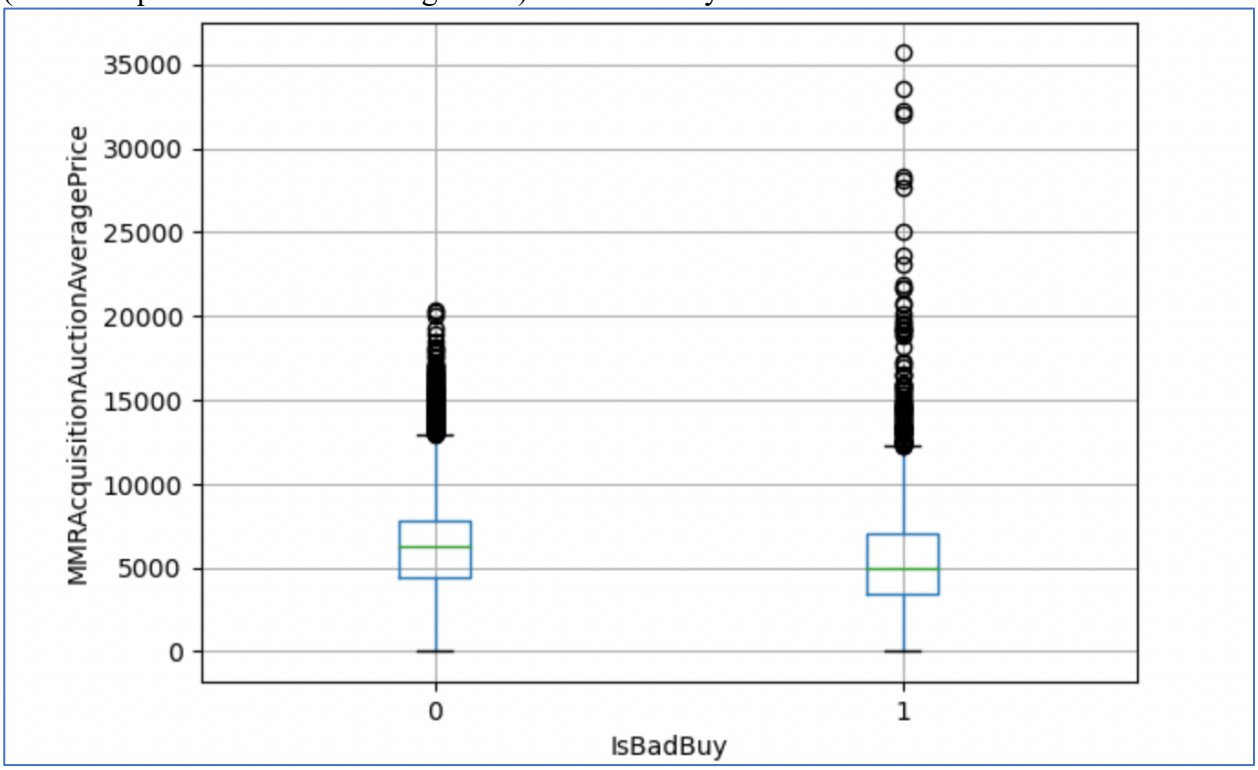
- Plotted a **heatmap** for all numerical columns, and then found that these are the factors which are impacting price of the vehicle:
    - **VehBCost** (Acquisition cost paid for the vehicle at time of the purchase): This factor has a correlation coefficient of more than 0.7 on all kinds of price parameters
    - **VehYear** (The manufacturer's year): In all the price types, vehicle year is having correlation coefficient of more than 0.5. **The latest the manufacturing of the vehicle, more will be the price**
    - **VehAge** (The years elapsed since the manufacturer's year): This variable has a negative impact on the price of the cars, more the age of the vehicle, less the price. That's why older cars sell for lesser price.
- PurchDate: From the time-series plot of purchase date and bad buys, it is found that some peaks are found on Sept 2009, Feb 2010, Jan 2011.



- **The proportion of bad buys is more in aged cars**, the proportion of bad buys is just 0.04 is cars of age 1, whereas this proportion is 0.3 in cars of age 9.

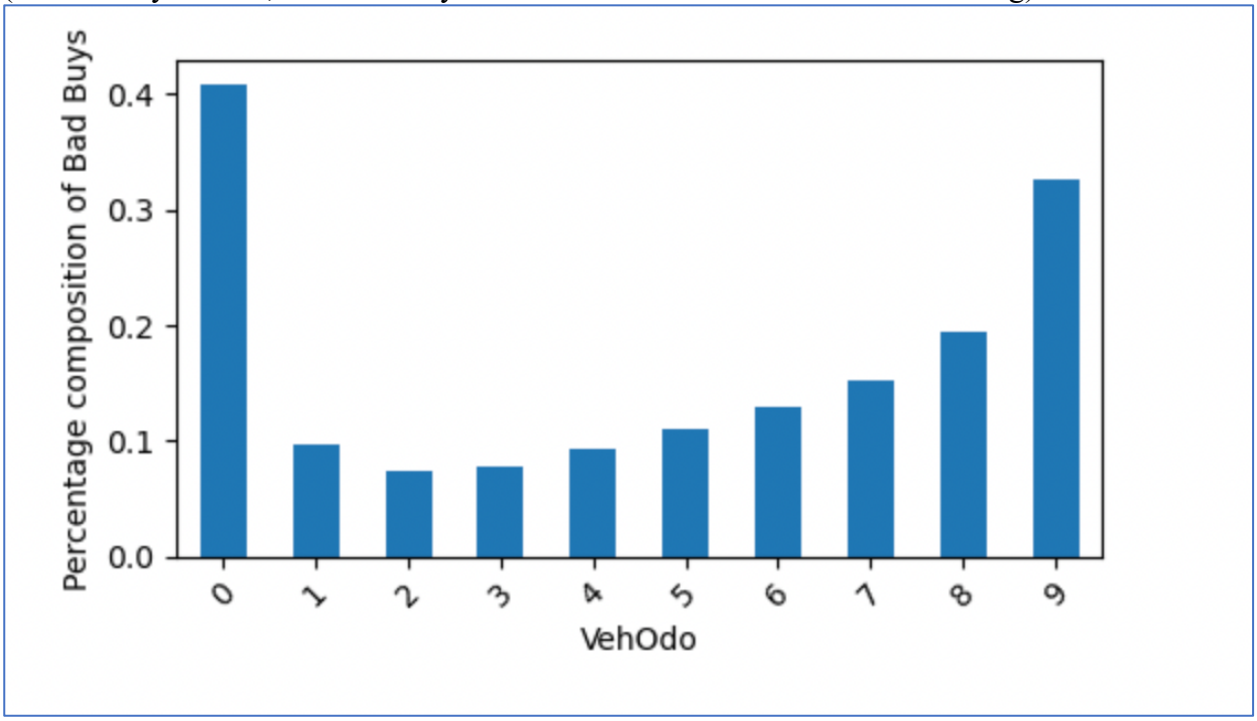


- **MMRAcquisitionAuctionAveragePrice** : The box plot supports our claim that MMRAcquisitionAuctionAveragePrice is a leading indicator for IsbadBuy. The median cost (MMRAcquisitionAuctionAveragePrice) of a bad buy is less than the median cost (MMRAcquisitionAuctionAveragePrice) of a Good buy



b. Summarize your recommendations for Carvana.

- **Pay extra surveillance/attention on cars with low odometer reading**. According to our analysis, even the cars with less odometer reading have a proportion of 0.4 of bad buys. (out of every 10 cars, 4 are bad buys in cases of cars with low odometer reading)

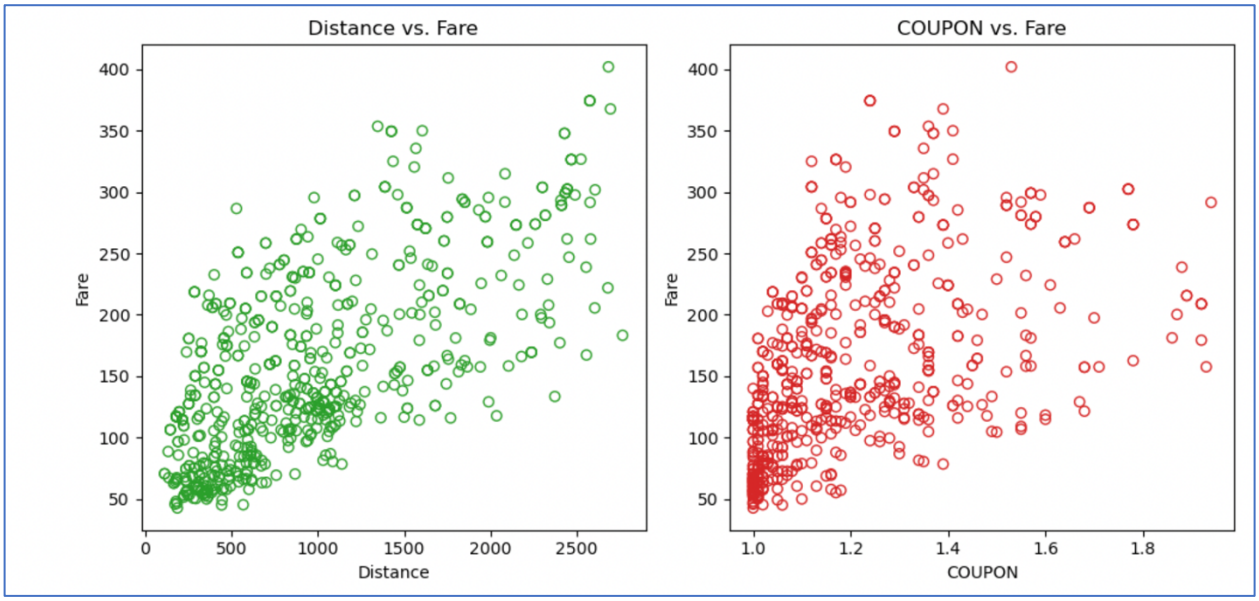


- **Do not bid higher for aged cars**, as there is a 30% chance that an aged car could be a bad buy.
- **From a domain expert, try to find the reasons of the peaks during Sept 2009, Feb 2010, Jan 2011**. Some external factors like supply chain of cars could be the potential reason

Use Airfares dataset

a. **Explore the relationship between FARE and other numerical predictors. Summarize your observations.**

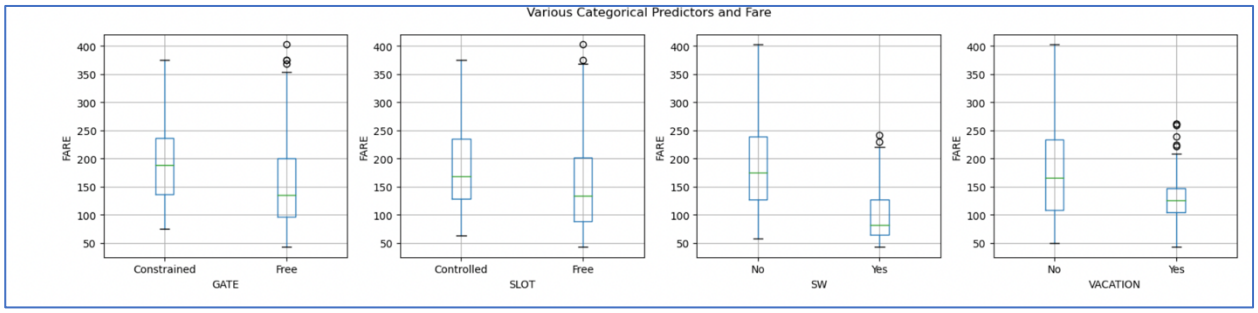
- **Strong Positive Correlation between FARE and DISTANCE (0.7**): A correlation of 0.7 indicates a strong positive linear relationship between FARE and DISTANCE. This means that as the distance of the travel increases, the fare tends to increase as well. This is an expected relationship, and it signifies that longer distances typically need more consumption of fuel making the fares go higher.
- **Moderate Positive Correlation between FARE and COUPON (0.5):** A correlation of 0.5 indicates a moderate positive relationship between the fare and the coupon variable. This suggests that as the coupon value increases, the fare will increase as well. This relationship might be due to discounts offered with higher coupon values.



- **Weak Correlation S_Income and E_POP:** These two variables are also positively related with Fare, but the correlation coefficient is only 0.3
- **Negative Correlation between FARE and PAX (-0.1**): A correlation of -0.1 is relatively weak, but it suggests a slight negative relationship between the number of passengers (PAX) and the fare. This might indicate that as the number of passengers increases, the fare per passenger may decrease. But this correlation is weak, so it's not a very strong indication of the former being true**.**

b. Use various tables to analyze the effect of categorical predictors on FARE.

1. The Variable **SW** is affecting the Fare very much, if SW is yes then the median fare is around 75, whereas if SW is No then the median Fare is 175
2. Whether it is **vacation** time or not, is also significantly impacting the fare price, if it is vacation then the median fare is around 125, Also there are outliers which is a common scenario when fares are high for the last-minute bookings.
3. Similarly, if the **gate** is FREE then the Fares are less than the minimum fare of a constrained gate. This may be because of the infrastructure cost of a constrained gate.
4. There is also a difference in the median fare of controlled slot v/s free slot. For controlled slots, the median fare is around 175, whereas for Free slot it is 140.



Thus, the four main categorical variables effecting Fare are Gate, Slot, SW and Vacation.

c. Develop a model to predict FARE. Summarize the accuracy measures based on the validation data. What is the final model you would recommend predicting FARE?

Solution: Based on our analysis till now, we are taking the following variables into our model

- Categorical: Gate, Slot, SW and Vacation
- Numerical/Continuous: Distance and Coupon

The test data size is 40%, we created a linear regression model, **the accuracy measures on the validation data are given as:**

- Mean Error (ME) : -1.1219
- Root Mean Squared Error (RMSE) : 40.8149
- Mean Absolute Error (MAE) : 32.2007
- Mean Percentage Error (MPE) : -7.284
- Mean Absolute Percentage Error (MAPE) : 24.0312

The wights of the predictors is given as:

- **intercept** 163.32286492035965

Predictor coefficient

1. DISTANCE 0.077489
2. COUPON -12.993141
3. GATE_Free -27.208234
4. SLOT_Free -15.072332
5. SW_Yes -52.554128
6. VACATION_Yes -49.925568

Thus, the final model equation would be given by:
Fare = (0.077489 * Distance) – (12.993141*Coupon) – (27.208234 * Gate_Free) – (15.072332 * Slot_Free) – (52.554128 * SW_Yes) – (49.925568 * Vacation_Yes)

d. Suppose a new airport is brought into service. Airlines have received news and are working on their prices. Would your model be helpful for them? Why or why not?
Solution: Yes, our model can help a new airport in the following ways:

- We have analyzed the past data and identified the important factors which impact Fare price. A replication of the model on the new data of the new airport can help them too
- We have given an equation identifying which factors are positively impacting the fare and which factors are negatively impacting the fare price. Thus is the airport and the airlines want to increase revenue fare then the following would help

Factors decreasing Revenue	Factors Increasing Revenue
Coupon	Distance (Long distance flights will bring more revenue)
Gate_Free	
Slot_Free	
SW_Yes	
Vacation_Yes	

Limitations: Since our model is based on the factors which were already available to us, thus any new factor if available to the new airport has not been taken into consideration. This new factor may or may not be significant.