**1. Carvana case:**

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

Solution:

Carvana Dataset consists of the following variables which are directly/indirectly supposed to impact the ‘IsBadBuy’ factor:

Continuous Numerical Columns:

['VehOdo', 'MMRAcquisitionAuctionAveragePrice', 'MMRAcquisitionAuctionCleanPrice', 'MMRAcquisitionRetailAveragePrice', 'MMRAcquisitonRetailCleanPrice', 'MMRCurrentAuctionAveragePrice', 'MMRCurrentAuctionCleanPrice', 'MMRCurrentRetailAveragePrice', 'MMRCurrentRetailCleanPrice', 'BYRNO', 'VNZIP1', 'VehBCost', 'WarrantyCost']

Low Cardinality Numerical Columns:

['VehYear', 'VehicleAge', 'WheelTypeID', 'IsOnlineSale']

Low Cardinality Categorical Columns:

['Auction', 'Transmission', 'WheelType', 'Nationality', 'TopThreeAmericanName']

High Cardinality Categorical Columns:

['Make', 'Model', 'Trim', 'SubModel', 'Color', 'Size', 'VNST']

**Leading Indicators:**

1. **Make** - Vehicle Manufacturer

A graph of blue bars

Description automatically generated

Major Manufacturers which contribute to ‘IsBadBuy’ are:   
**Plymouth, Lexus, Infiniti, Mini, Lincoln, Acura**

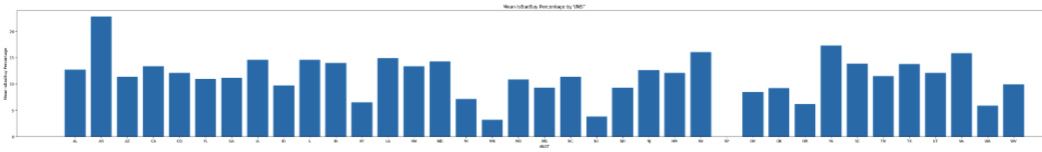
2. **Trim** - Vehicle Trim Level

A graph of blue lines

Description automatically generated with medium confidence

Major Trim levels which contribute to ‘IsBadBuy’ are: **JLX, Maz, Out, Z24, GXP**

1. **VNST** - State where the the car was purchased



Major states which contribute to ‘IsBadBuy’ are: **AR, PA, NV, VA**

1. **Auction** - Auction provider at which the vehicle was purchased

A graph with a line

Description automatically generated

Major Contributing Auction to ‘IsBadBuy: **ADESA**

5. WheelType - The vehicle wheel type description

A graph with a line

Description automatically generated with medium confidence

Major Wheel Type which contribute to ‘IsBadBuy’ are: Special, Alloy

1. **Nationality**

A graph on a white background

Description automatically generated

Major Nationalities which contribute to ‘IsBadBuy’ are: Other, Other Asian, Top Line Asian

1. **TopThreeAmericanName** - Identifies if the manufacturer is one of the top three American manufacturers A graph showing a line

   Description automatically generated

Out of three major American Manufacturers: Ford is contributing to ‘IsBadBuy’ the most

1. **VehYear** - The manufacturer's year of the vehicle

A graph of blue squares

Description automatically generated

With the increase in year the ‘IsBadBuy’ decreases, Inversely related

1. **VehicleAge** - The Years elapsed since the manufacturer's year

A graph with blue squares

Description automatically generated

With the increase in car’s age the ‘IsBadBuy’ increases, Directly Related

1. WheelTypeID - The type id of the vehicle wheel A graph with blue squares

   Description automatically generated

As shown in the figure above, the wheelTypeID 0.0 is contributing maximum to the ‘IsBadBuy’

1. IsOnlineSale - Identifies if the vehicle was originally purchased online

A graph with a bar and a number of text

Description automatically generated with medium confidence

As shown in the figure, Online Sale is not affecting much to the ‘IsBadBuy’

1. **VehOdo** - The vehicles odometer reading

A graph with lines and dots

Description automatically generated

**Median Odometer Reading of ‘IsBadBuy’ vehicles are higher**

1. VehBCost - Acquisition cost paid for the vehicle at time of purchase

A graph with black and orange lines and dots

Description automatically generated

Higher Acquisition cost of majority of ‘IsBadBuy’ labeled vehicles

1. WarrantyCost - Warranty price (term=36month and millage=36K)

A graph of a number of boxes

Description automatically generated with medium confidence

1. MMRAcquisitionAuctionAveragePrice - Acquisition price for this vehicle in average condition at time of purchase

MMRAcquisitionAuctionCleanPrice - Acquisition price for this vehicle in the above Average condition at time of purchase

MMRAcquisitionRetailAveragePrice - Acquisition price for this vehicle in the retail market in average condition at time of purchase

MMRAcquisitonRetailCleanPrice - Acquisition price for this vehicle in the retail market in above average condition at time of purchase

A group of graphs with numbers and lines

Description automatically generated with medium confidence

MMRCurrentAuctionAveragePrice - Acquisition price for this vehicle in average condition as of current day

MMRCurrentAuctionCleanPrice - Acquisition price for this vehicle in the above condition as of current day

MMRCurrentRetailAveragePrice - Acquisition price for this vehicle in the retail market in average condition as of current day

MMRCurrentRetailCleanPrice - Acquisition price for this vehicle in the retail market in above average condition as of current day A group of graphs with numbers and lines

Description automatically generated with medium confidence

1. Profit/Loss Vs ‘IsBadBuy’ (Extra) – Showing Profits are inversely proportional to ‘IsBadBuy’ value

A graph on a white background

Description automatically generated

**b. Summarize your recommendations for Carvana.**

Based on the visual Analytics in the first part of the question, we have the following recommendations for carvana:

1. Avoid the following car makers: **Plymouth, Lexus, Infiniti, Mini, Lincoln, Acura**
2. Avoid vehicles with Trim levels: **JLX, Maz, Out, Z24, GXP**
3. States to avoid: **AR, PA, NV, VA**
4. Prefer Auction ‘Manheim’ over ‘**Adesa’**
5. **Avoid vehicles with ‘Special’ Wheeltype**
6. ‘**American’** cars are the performing best in terms of ‘IsBadBuy’
7. Among the top Three American Names, **buying ‘Ford’ cars are leading to higher ‘IsBadBuy’ values**
8. Cars with **higher vehicle year** should be preferred
9. Cars with **lower Vehicle age** should be preferred
10. **WheelType 0.0 should be avoided**
11. **Lower Odometer Reading** should be preferred
12. **Lower Acquisition Costing** vehicles can give more profits, lower ‘IsBadBuy’ value,
13. ‘IsBadBuy’ values are highly getting Influeced by MMR Prices, **higher MMR prices are being shown by ‘IsBadBuy’ labeled vehicles.**

2. Use Airfares dataset. You may ignore the first 4 variables.  
**Airfares Analysis** You may ignore the first 4 variables.

**Data Dictionary Airfares**

~~1.       S\_CODE: starting airport’s code~~

~~2.       S\_CITY: starting city~~

~~3.       E\_CODE: ending airport’s code~~

~~4.       E\_CITY: ending city~~

5.       COUPON: average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon

          flight is a one stop flight, etc.) for that route

6.       NEW: number of new carriers entering that route between Q3-96 and Q2-97

7.       VACATION: whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are

          generally considered vacation routes

8.       SW: whether Southwest Airlines serves that route (Yes) or not (No)

9.       HI: Herfindel Index – measure of market concentration (refer to BMGT 681)

10.     S\_INCOME: starting city’s average personal income

11.     E\_INCOME: ending city’s average personal income

12.     S\_POP: starting city’s population

13.     E\_POP: ending city’s population

14.     SLOT: whether either endpoint airport is slot controlled or not; this is a measure of airport

          congestion

15.     GATE: whether either endpoint airport has gate constraints or not; this is another measure of airport congestion

16.     DISTANCE: distance between two endpoint airports in miles

17.     PAX: number of passengers on that route during period of data collection

18.     FARE: average fare on that route

**Q2.1: Explore the relationship between FARE and other numerical predictors. Summarize your observations.**

*Table 1.1 Summary statistics for this data set in numerical variables*

A table of numbers and numbers

Description automatically generated

*Table 1.2 Heatmap for all Variables*

A diagram of a number of squares

Description automatically generated with medium confidence

*Table 1.3*

A group of graphs showing different numbers

Description automatically generated with medium confidence

*Table 1.3 shows scatterplots for each numerical variable by FARE. We do see a correlation between FARE and the following variables:*

* COUPON - slightly positive correlation with this variable. Customers are more likely to go for higher fares when they have coupons.
* DISTANCE - there is a correlation between DISTANCE and FARE.  The longer the distance traveled the more fuel consumed and thus the airlines charge higher prices.

We do not see any significant correlation between E\_INCOME and S\_INCOME (if anything it is slightly there) and PAX (number of passengers).

**Q2.2:   
Use various tables to analyze the effect of categorical predictors on FARE.**

A graph of different sizes and shapes

Description automatically generated with medium confidence

There are four categorical variables:

1. VACATION - Vacation routes have more vacation travelers who are looking for a good deal(s) vs business travelers who are looking for convenience and ease of travel and are willing to pay extra
2. SW - Southwest is a budget carrier and having Southwest on that route means you are competing with Southwest which has lower price(s) and fares on the route will be lower to compete.
3. SLOT -  - Slot measures congestion and Controlled gates have more congestion and show a higher average for FARE.
4. GATE - Constrained Gate has more congestion and thus has a higher average FARE

**Q2.3:   
Develop a model to predict FARE. Summarize the accuracy measures based on the validation data. What is the final model you would recommend to predict FARE?**

A screenshot of a document

Description automatically generated

* The accuracy can be assessed by looking at RMSE which is slightly positive.
* Reducing the number of variables does not improve the accuracy - it in fact reduces the accuracy as evidenced by high values for RMSE, MAE …
* The values for SLOT and GATE, which are a marker of congestion (and thus higher demand) are positively correlated with higher fares.  Additionally, COUPONS, lead to higher fares.

A screenshot of a computer

Description automatically generated

**Q2.4:**

**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?**

* Our model primarily predicts airfare (FARE) based on factors such as coupon availability, income levels, population statistics, and distance. In our model, categorical predictors like Vacation travel, SW (stopover), Gate, and time slot of departure and arrival have proven to be significant influencing factors. Simultaneously, among the numerical factors, Coupon and New are among the important contributors.
* When a new airport is introduced, airlines will undoubtedly increase the number of flights to accommodate the increased supply. This can lead to a potential decrease in prices if the demand doesn't rise proportionally, as our model suggests. However, this is a fundamental assumption in our model.
* Airfare pricing is influenced by various dynamic and real-time factors, such as fluctuations in fuel prices, pricing strategies of competitors, seasonal variations in demand, and more.
* Therefore, while our model provides a foundational pricing strategy for airlines, in practice, airlines need to closely monitor market dynamics and adjust prices based on real-time information. The model can serve as a valuable reference tool but cannot replace the need for keen insights into market trends and competitive environments.
* In summary, our model provides airlines with a robust analytical tool, but in practice, flexibility and market acumen are equally important.