Harshaanth Thiyagaraja Kumar

Several new airports have opened in major cities, opening the market for new routes (a route refers to a pair of airports), and Southwest has not announced whether it will cover routes to/from these cities. In order to price flights on these routes, a major airline collected information on 638 air routes in the United States. Some factors are known about these new routes: the distance traveled, demographics of the city where the new airport is located, and whether this city is a vacation destiny. Other factors are yet unknown (e.g., the number of passengers that will travel this route). A major unknown factor is whether Southwest or another discount airline will travel on these new routes. Southwest's strategy (point-to-point routes covering only major cities, use of secondary airports, standardized fleet, low fares) has been very different from the model followed by the older and bigger airlines (hub-and-spoke model extending to even smaller cities, presence in primary airports, variety in fleet, pursuit of high-end business travelers). The presence of discount airlines is therefore believed to reduce the fares greatly.

The file Airfares_hw1 contains real data that were collected for the third quarter of 1996.

1. Read the problem description carefully, and choose the dependent variable

fare

2. Based on your knowledge, which independent variables would you use in the dataset? Explain.

For example, the distance between two airports may impact fare, it's possible that fare would increase when the distance increases.

3. Any thoughts of variables that could be collected outside this dataset?

For example, season may impact fare. When it's busy season, the fare would be higher.

4. Which variables are numerical? Which variables are categorical?

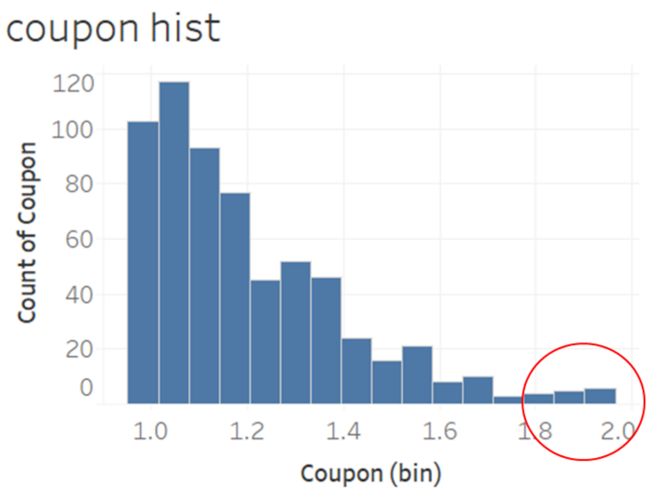
Numerical: new, distance, pax, coupon

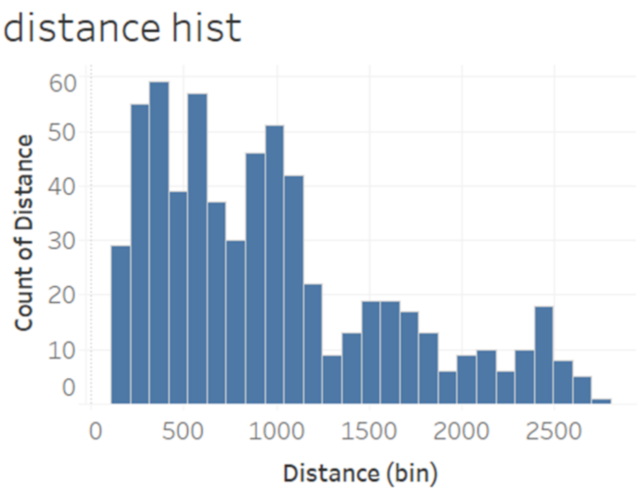
Categorical: vacation, sw, slot, gate

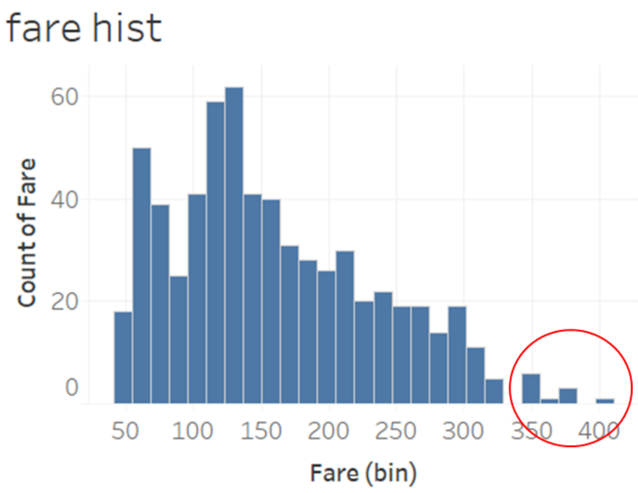
5. Explore the dataset, any missing values? If yes, please deal with the missing values.

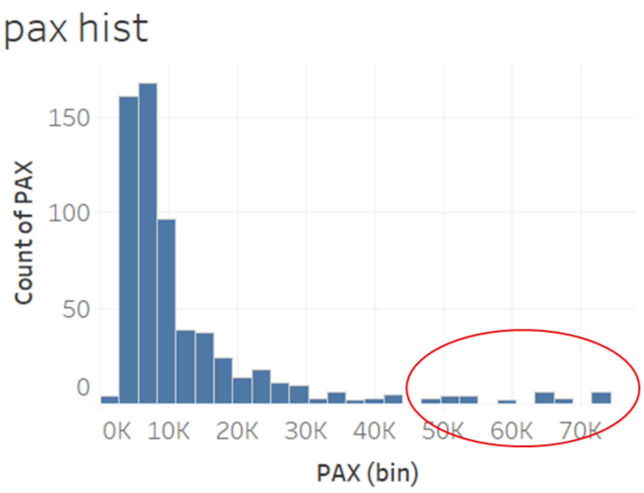
Yes, there are missing values in sw, distance, coupon and fare. Considering we only have a few missing values, we could directly delete those rows.

6. Using Tableau to create histograms of all numerical variables, any outliers? If yes, please circle them on your screenshots. (paste screenshots)


coupon hist

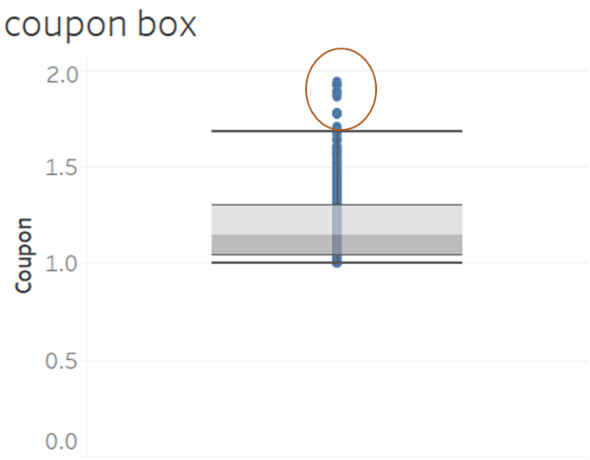

distance hist

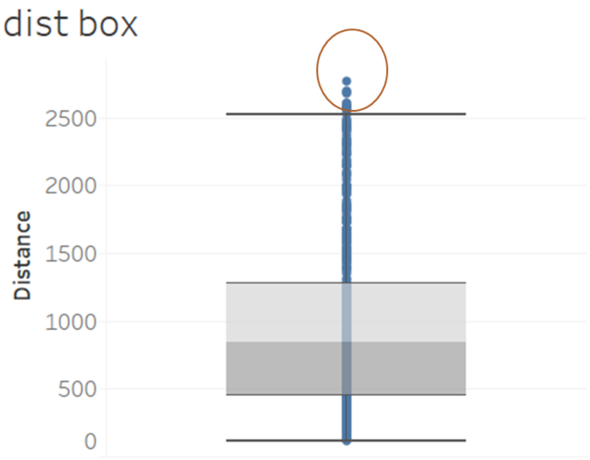

fare hist

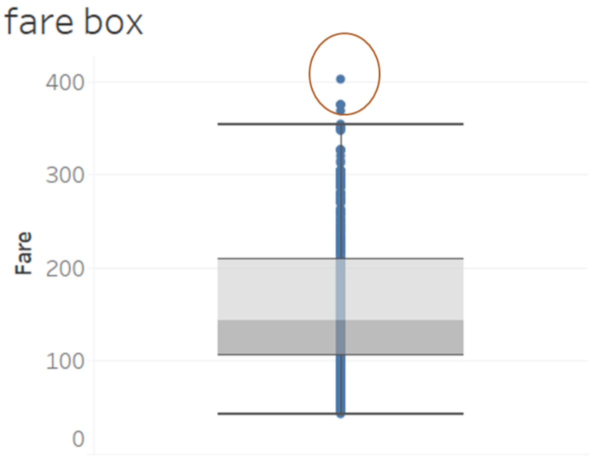

pax hist

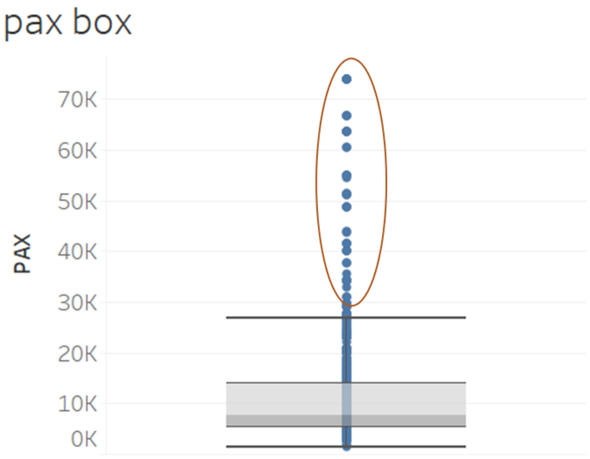
Circles of outliers are not necessarily to be exactly the same as they showed in the picture. It's just an estimated range showing that those could be outliers.

7. Using Tableau to create boxplots of all numerical variables, any outliers? If yes, please circle them on your screenshots. (paste screenshots)





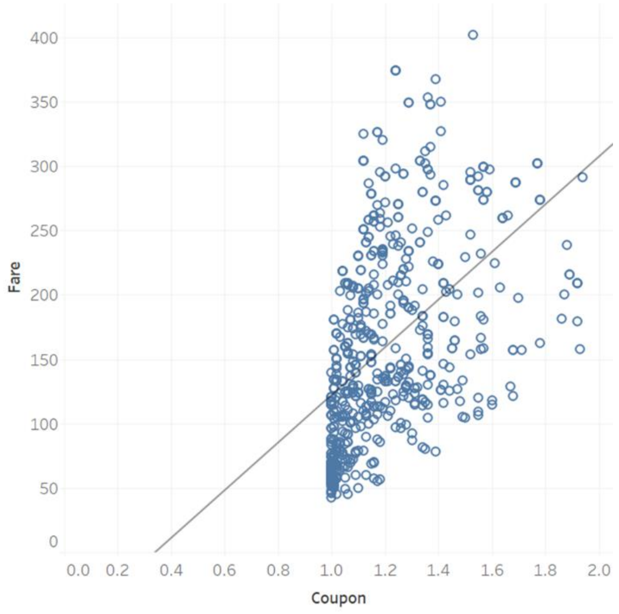




8. Using Tableau to create the scatter plots of fare and all the other numerical variables and show the trend line, any trend? If yes, please explain the trend. (paste screenshots)

Harshaanth Thiyagaraja Kumar

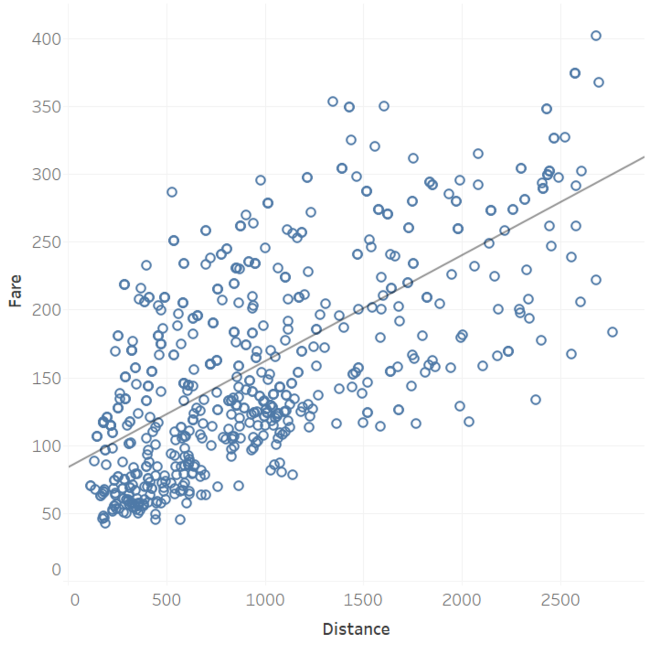
fare and coupon



Coupon vs. Fare.

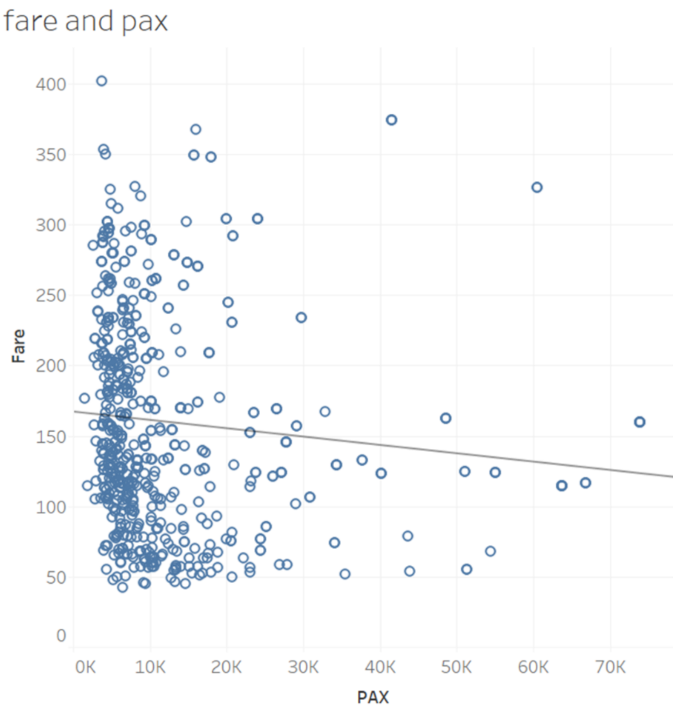
Fare and coupon are positively correlated. When there are more stops, fare is higher.

fare and dist



Distance vs. Fare.

Distance and fare are positively correlated. When the distance between two airports increase, fare increases.



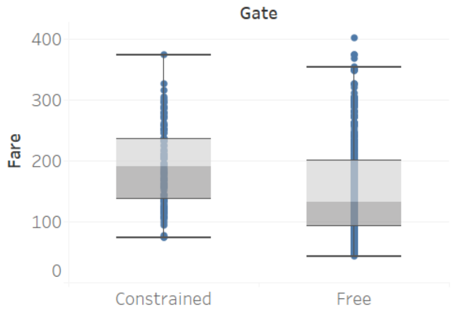
PAX vs. Fare.

Fare and the number of passengers on the route are negatively correlated. When there are more passengers, fare would be lower.

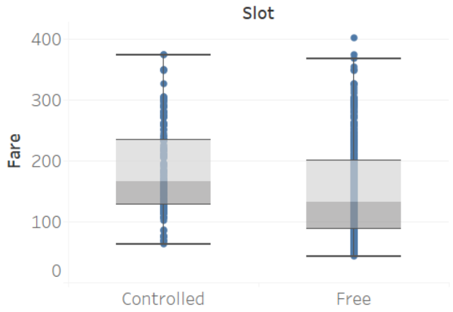
9. Using Tableau to create the box plots of fare and all the other categorical variables, any observations? If yes, please explain. (paste screenshots)

Harshaanth Thiyagaraja Kumar

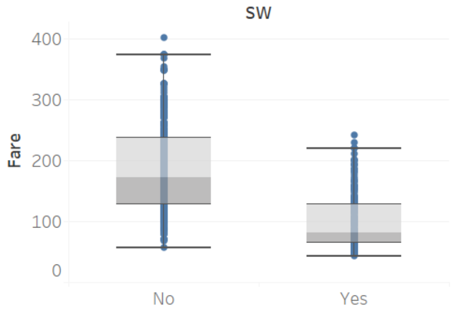
fare and gate



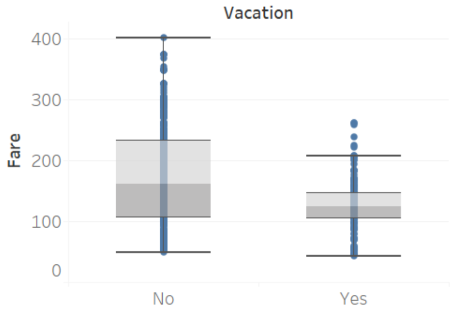
fare and slot



fare and sw



fare and vacation



Fare is higher when there are airport congestions.

If southwest airlines serves the route, fare is lower.

If the route is a vacation route, fare is lower.