Assessment 1

Exploratory Data Analysis and Visualization

Ziyang Lin

Jan 31, 2022

Question 1

(a). As n increases (i.e. simulate more data points from the proposed distribution), we would have more data to compare with the emprirical distribution of our data set, and any unusual visual pattern in the Q-Q plot can be easier to detect. Ideally, we would like n to at least equal to m so that each quantile can be matched in the Q-Q plot. In general, the higher n is, the more confident our conclusion would be regarding the comparison of the two distributions.

(b). In this Q-Q plot, we see on the points align with the straight line only around the centre. There are two main observations. In the lower end, the points deviate from the straight line towards the bottom, while in the upper end, the points deviate from the straight line towards the top. If we only observe the first observation, we can say the data appears to be left-skewed. Similarly, if we only observe the second, we can say the data appears to be right-skewed. However, in this case we see heavy-tails on both ends, this means we have positive excess kurtosis in the data as kurtosis is the measure of tailness of the distribution. There is another possibility that the data is actually bimodal to have this Q-Q plot.

Question 2

We first load the data set into R environment and see the below first few rows.

```r
travels <- read.csv("~/Desktop/travel-times.csv")
knitr::kable(head(travels), caption="First Few Rows of Dataset")
```

Table 1: First Few Rows of Dataset

Date	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	TotalTime
01/06/2012	Friday	Home	51.29	127.4	78.3	39.3
01/06/2012	Friday	Work	51.63	130.3	81.8	37.9
01/04/2012	Wednesday	Home	51.27	127.4	82.0	37.5
01/04/2012	Wednesday	Work	49.17	132.3	74.2	39.8
01/03/2012	Tuesday	Home	51.15	136.2	NA	36.8
01/03/2012	Tuesday	Work	51.80	135.8	84.5	36.8

(a). From the above table, we see there are 7 variables for each observation.

- `Date`: it can be considered *nominal*, *ordinal*, or *interval* type. If the context states that only equality of multiple `Date` can be compared but not the difference, then it is *nominal*. If we consider the time order in it (as it is a time indicator), then it can be *ordinal*. If the difference between any two `Date` values are also meaningful in the context, this would make it an *interval* type.
- `DayOfWeek`: categorical variable with 5 levels, so a *nominal* type. If we consider the order (Monday as

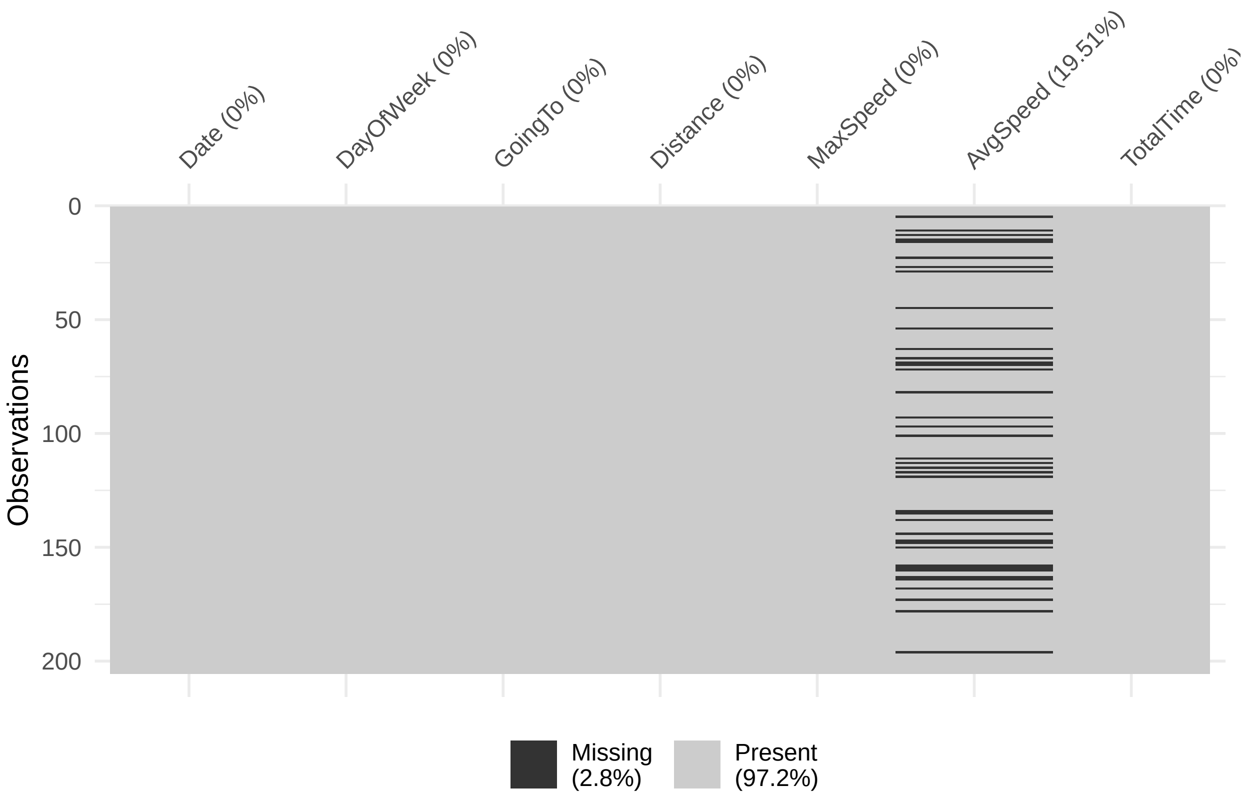
earliest, and Friday as latest), then it can also be *ordinal* type.

- `GoingTo`: categorical variable with 2 levels (i.e. binary variable), so a *nominal* type.
- `Distance`: the "zero" value of `Distance` is meaningful, and that it can support all mathematical operations, this makes it a *ratio* type.
- `MaxSpeed`: same as `Distance`, a *ratio* type.
- `AvgSpeed`: same as above, a *ratio* type.
- `TotalTime`: same as above, a *ratio* type.

(b). If `AvgSpeed` is not MNAR, then it is either MAR or MCAR. We can plot `AvgSpeed` against either or both of the two time variable to see whether its missingness is conditioned on these variables.

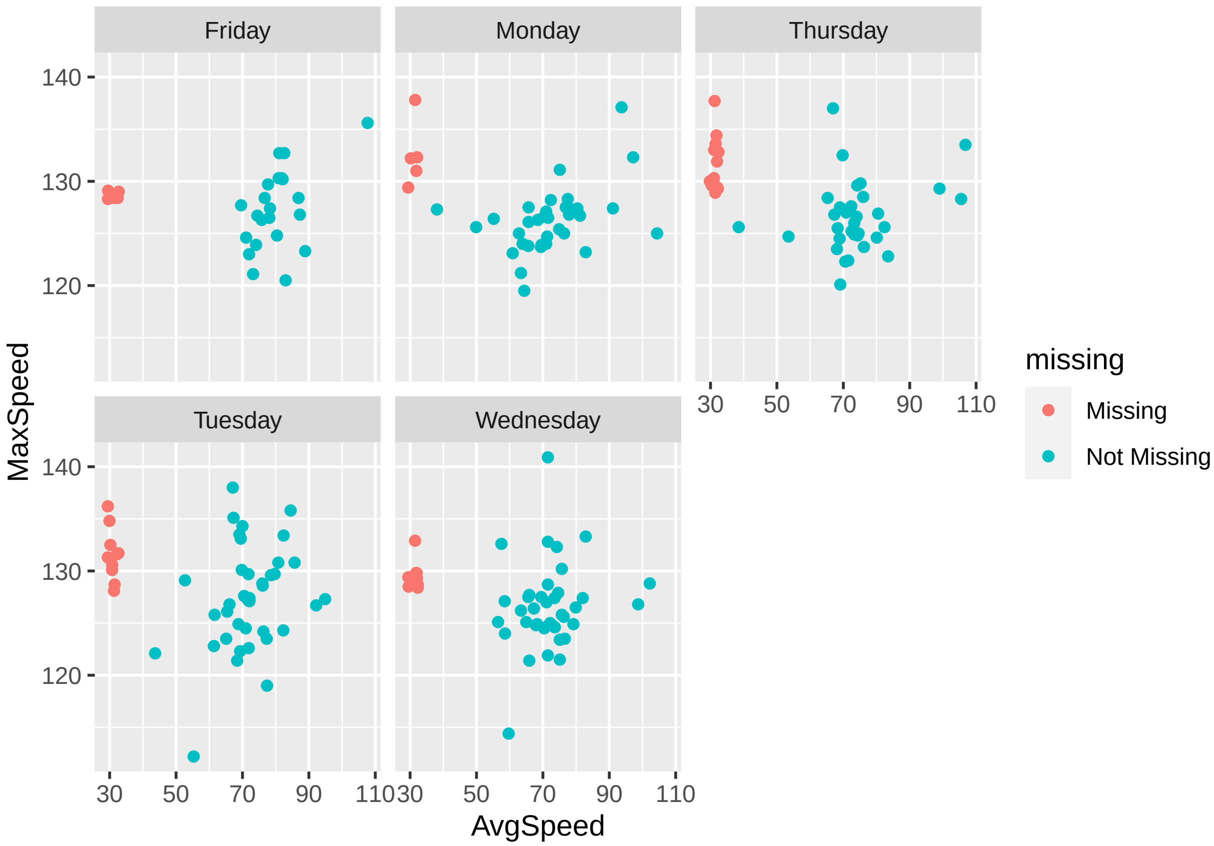
We first consider the missing locations in all observations. We see that the distribution of missing data looks random and evenly spread across all observations.

```
vis_miss(travels)
```



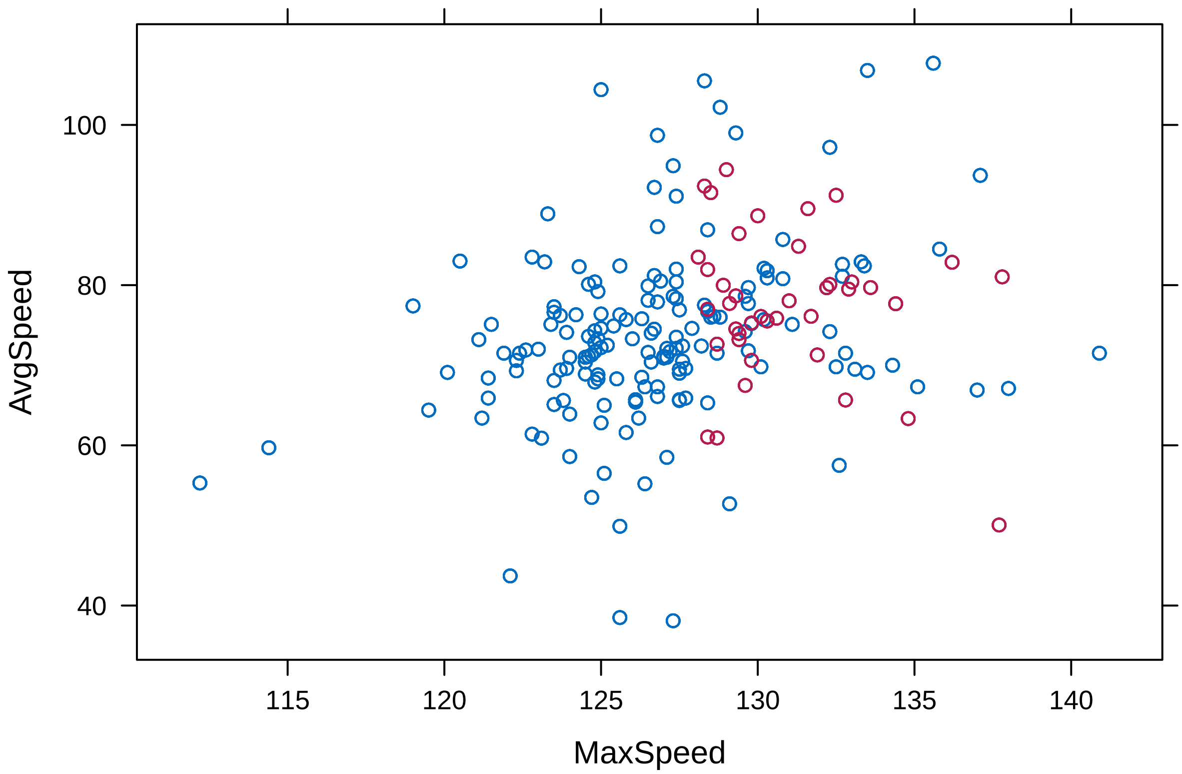
Next we condition on `DayOfWeek`. We see that all five days have some missing records. Though Monday and Friday appear to have less, but these two days also have fewer records, so overall the difference is negligible, so we can conclude that the missingness of `AvgSpeed` does not depend on `DayOfWeek`. A similar plot can be generated with `Date` (treating it as categorical), and we also detect no unusual pattern. This means we can categorize the missingness as MCAR.

```
ggplot(travels,
       aes(x=AvgSpeed, y=MaxSpeed)) +
  geom_miss_point() +
  facet_wrap(~DayOfWeek)
```



(c). We have determined that `AvgSpeed` is MCAR, this means both mean imputation and regression imputation will produce unbiased mean estimate. We will deploy regression imputation here and see below graph for the imputed data (red dots.)

```
travels2 <- travels[,!names(travels) %in% c("Date")]
impTravels <- mice(travels2, seed=2, method="norm.predict", m=10, maxit=5, print=FALSE)
xyplot(impTravels, AvgSpeed~MaxSpeed)
```



We then modify the original data set to include these imputed data, and consider the below summary table (without `Date`), we see no data is missing anymore.

```
travels_lr <- complete(impTravels)
travels_lr$Date <- travels$Date
knitr::kable(summary(travels_lr)[,1:6], caption="Summary After Imputation")
```

Table 2: Summary After Imputation

DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	TotalTime
Friday :27	Home:100	Min. :48.32	Min. :112.2	Min. : 38.10	Min. :28.2
Monday :39	Work:105	1st Qu.:50.65	1st Qu.:124.9	1st Qu.: 69.00	1st Qu.:38.4
Thursday :44	NA	Median :51.14	Median :127.4	Median : 74.00	Median :41.3
Tuesday :48	NA	Mean :50.98	Mean :127.6	Mean : 74.34	Mean :41.9
Wednesday:47	NA	3rd Qu.:51.63	3rd Qu.:129.8	3rd Qu.: 79.90	3rd Qu.:44.4
NA	NA	Max. :60.32	Max. :140.9	Max. :107.70	Max. :82.3
NA	NA	NA	NA	NA	NA

(d). The modified Z-score for outlier detection aims for data points with $|M_i| > 3.5$ where:

$$M_i = \frac{0.6745(x_i - \bar{x})}{MAD}, \text{ where } MAD = \text{median}_i\{|x_i - \bar{x}|\}$$

Consider the below code to compute modified Z-score M_i for each data points

```
med <- median(travels$TotalTime)
MAD <- mad(travels$TotalTime)
m <- 0.6475 * (travels$TotalTime - med) / MAD
head(sort(m, decreasing=TRUE)) # top positive values
```

```
## [1] 5.776143 4.113741 3.409333 2.803543 2.099135 2.085047
```

```r
head(sort(m)) # top negative values
```

```
## [1] -1.845548 -1.831460 -1.803284 -1.761019 -1.718755 -1.676490
```

We see that there are two points being above the $|M_i| = 3.5$ threshold. Now we create a tibble that contains only these two points.

```r
(out <- travels %>%
   mutate(ModZScore=abs(0.6475 * (TotalTime - median(TotalTime)) / MAD)) %>%
   filter(ModZScore > 3.5))
```

```
##         Date DayOfWeek GoingTo Distance MaxSpeed AvgSpeed TotalTime ModZScore
## 1 11/21/2011    Monday    Work    52.25    127.3     38.1      82.3  5.776143
## 2  7/26/2011   Tuesday    Home    51.28    122.1     43.7      70.5  4.113741
```