hw1

Ishaan Dey

2/4/2021

1. Import

```
library(haven)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(purrr)
library(cowplot)
```

```
flights <- read_dta("version10_RA.dta") %>% tibble()
```

2. Data Definitions

– market - the combination of origin and destination airports

– origin - the origin airport

– dest - the destination airport

– year - year when the travel occurred

– quarter - quarter when the travel occurred

– carrier - the carrier who transported the passengers (e.g. American – AA)

– nonstopmiles mkt - nonstop distance in miles

– totalpassengers - passengers transported by the carrier

– medianmktfare mkt - median fare charged by the carrier

– meanmktfare mkt - mean fare charged by the carrier

3.

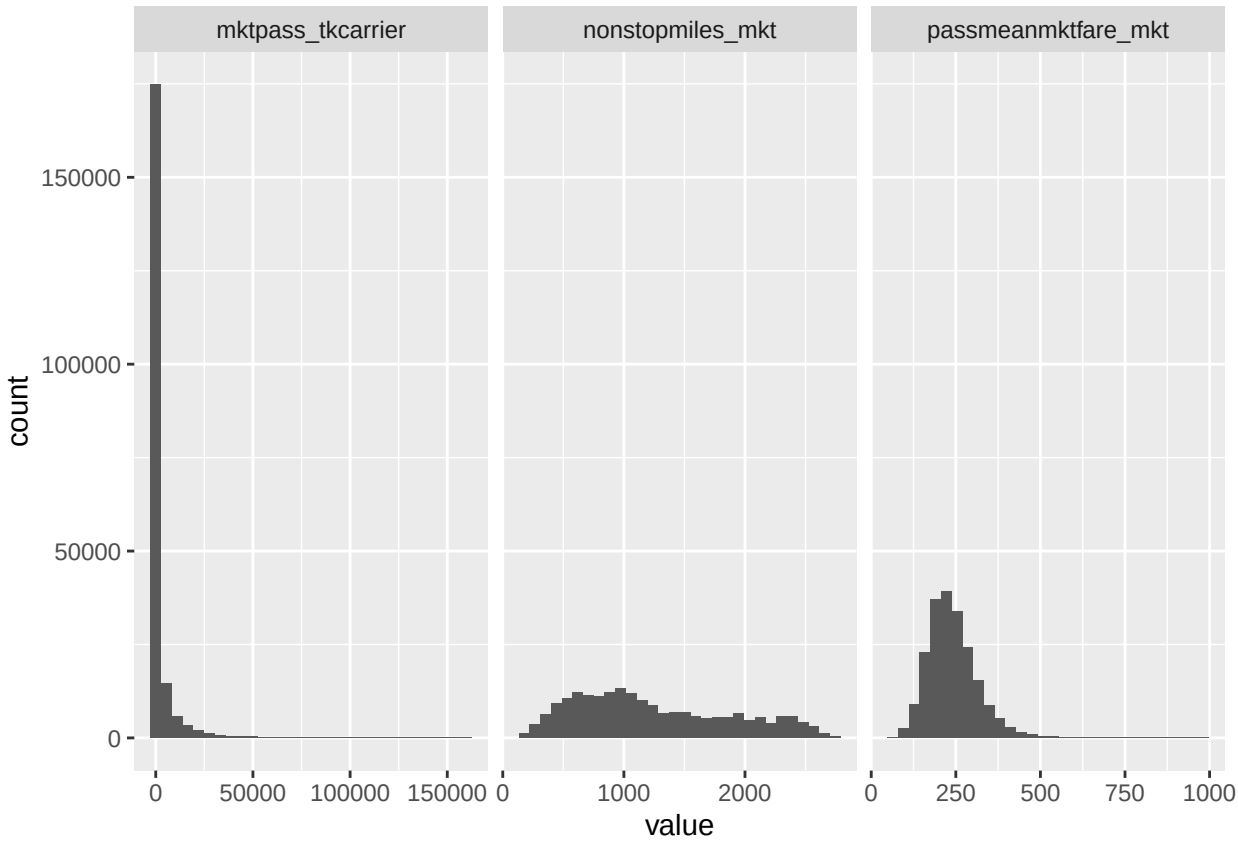
```
summary.stats <- function(x,...){
  c(mean=mean(x, ...),
  sd=sd(x, ...),
  min=min(x, ...),
  lower = quantile(x, ..., 1/4),
  median=median(x, ...),
  upper = quantile(x, ..., 3/4),
  max=max(x,...))
}

num.cols <- c("passmeanmktfare_mkt","mktpass_tkcarrier","nonstopmiles_mkt")

flights.summary <- sapply(select(flights, all_of(num.cols)), summary.stats) %>% as.data.frame()
flights.summary
```

```
##           passmeanmktfare_mkt mktpass_tkcarrier nonstopmiles_mkt
## mean                239.77945          2491.877        1242.8215
## sd                   72.96611          7120.953         637.9157
## min                  59.97292            90.000         151.7254
## lower.25%           189.53795           210.000         726.0604
## median              230.44464           480.000        1102.4904
## upper.75%           279.65482          1280.000        1733.1265
## max                 979.01361        160100.000        2719.2703
```

We can see that for `passmeanmktfare_mkt` and `mktpass_tkcarrier`, there are many outliers that bring up the maximum significantly over the median. From the histograms, we see that the distribution of `nonstopmiles_mkt` is fairly even, and confirms lack of outliers.

```
flights %>%
  select(all_of(num.cols)) %>%
  gather() %>%
  ggplot(aes(value)) +
  geom_histogram() +
  facet_grid(~key, scales = 'free_x')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



4

We'll first map all the carriers to a group, and check its distribution

```
carrier_groups = list(AA='Legacy',
                      DL='Legacy',
                      UA='Legacy',
                      WN='LCC',
                      B6='LCC')

flights$carrier_group = flights$tkcarrier %>% recode(!!!carrier_groups, .default='Other')
flights$carrier_group %>% table()
```

```
## .
##    LCC Legacy  Other
##  23887 107672  73214
```

```
group_stats =
for (g in flights$carrier_group %>% unique()){
  print(paste0('Carrier Group: ',g))
  group.stats <- sapply(flights %>%
           filter(carrier_group == g) %>%
           select(all_of(num.cols)),
         summary.stats) %>%
```

```
    as.data.frame()
  print(group.stats)
}
```

```
## [1] "Carrier Group: Legacy"
##           passmeanmktfare_mkt mktpass_tkcarrier nonstopmiles_mkt
## mean                254.82647          1957.027        1240.8139
## sd                   74.09149          6117.566         630.3758
## min                  68.07591            90.000         152.0161
## lower.25%           203.06381           200.000         731.0565
## median              245.52037           440.000        1101.3259
## upper.75%           294.55953          1080.000        1723.6611
## max                 923.81018        146590.000        2719.2703
## [1] "Carrier Group: Other"
##           passmeanmktfare_mkt mktpass_tkcarrier nonstopmiles_mkt
## mean                228.54031          2090.861        1247.4475
## sd                   71.72090          5671.783         642.5461
## min                  59.97292            90.000         151.7254
## lower.25%           178.99164           190.000         726.7436
## median              217.61833           420.000        1101.7332
## upper.75%           268.78526          1090.000        1741.6509
## max                 979.01361        114960.000        2719.2703
## [1] "Carrier Group: LCC"
##           passmeanmktfare_mkt mktpass_tkcarrier nonstopmiles_mkt
## mean                206.40230          6131.865        1237.6926
## sd                   51.07686         12344.255         657.1104
## min                  68.82486            90.000         153.2130
## lower.25%           171.48658           480.000         693.3195
## median              206.88779          1310.000        1111.8335
## upper.75%           241.41648          5830.000        1745.8091
## max                 462.05280        160100.000        2714.5547
```