R Notebook

```
library(readxl)
df_fares <- read_excel("data_sets/2022 Fare Revenue.xlsx")
df_funding <- read_excel("data_sets/2022 Federal Funding Allocation_1-2.xlsx")
```

drop columns Parent ID, Reporting type, TOS

```
# renaming the colnumns

colnames(df_fares) <- gsub(" ", "_", colnames(df_fares)) # replaced space with underscore between two w

colnames(df_fares)[1] <- "Parent_ID"

# removing columns which are not required

fare_revenue<- df_fares %>%
  select(-c(Parent_ID, Reporter_Type, TOS))

# changing column names to all lower

colnames(fare_revenue) <- tolower(colnames(fare_revenue))
```

```
# replacing NA with 0 for Org_paid_fair, when mode is DR

fare_revenue <- fare_revenue %>%
  mutate(organization_paid_fares = ifelse(is.na(organization_paid_fares), 0, organization_paid_fares), p
```

there are 18 mode of transportation in this data set, not all states have all the modes of transport availabel.

```
# checking for duplicate data

  sum(duplicated(fare_revenue))
```

```
## [1] 62
```

```
# filtering those duplicated values to verify

fare_revenue %>%
  filter(duplicated(fare_revenue)==TRUE)
```

```
## # A tibble: 62 x 8
##    ntd_id agency_name   reporting_module mode  expense_type passenger_paid_fares
##     <dbl> <chr>         <chr>            <chr> <chr>                       <dbl>
##  1     43 Chelan Dougl~ Urban            DR    Funds Earne~                    0
##  2    307 Coos County ~ Rural            DR    Funds Expen~                    0
##  3    309 Grant County~ Rural            DR    Funds Expen~                    0
##  4    309 Grant County~ Rural            DR    Funds Expen~                    0
##  5    376 Ride Connect~ Urban            DR    Funds Earne~                    0
##  6    378 Central Area~ Rural            DR    Funds Expen~                    0
##  7  10014 Worcester Re~ Urban            DR    Funds Earne~                    0
```

```
##  8  10014 Worcester Re~ Urban            DR    Funds Earne~                    0
##  9  10137 Advance Tran~ Rural            DR    Funds Expen~                    0
## 10  10137 Advance Tran~ Rural            DR    Funds Expen~                    0
## # i 52 more rows
## # i 2 more variables: organization_paid_fares <dbl>, total_fares <dbl>
```

```
# duplicated valeus can be safely removed, as removing them wouldn't really effect the outcome


df_clean_1 <- fare_revenue %>%
  filter(!duplicated(fare_revenue))

sum(duplicated(df_clean_1))
```

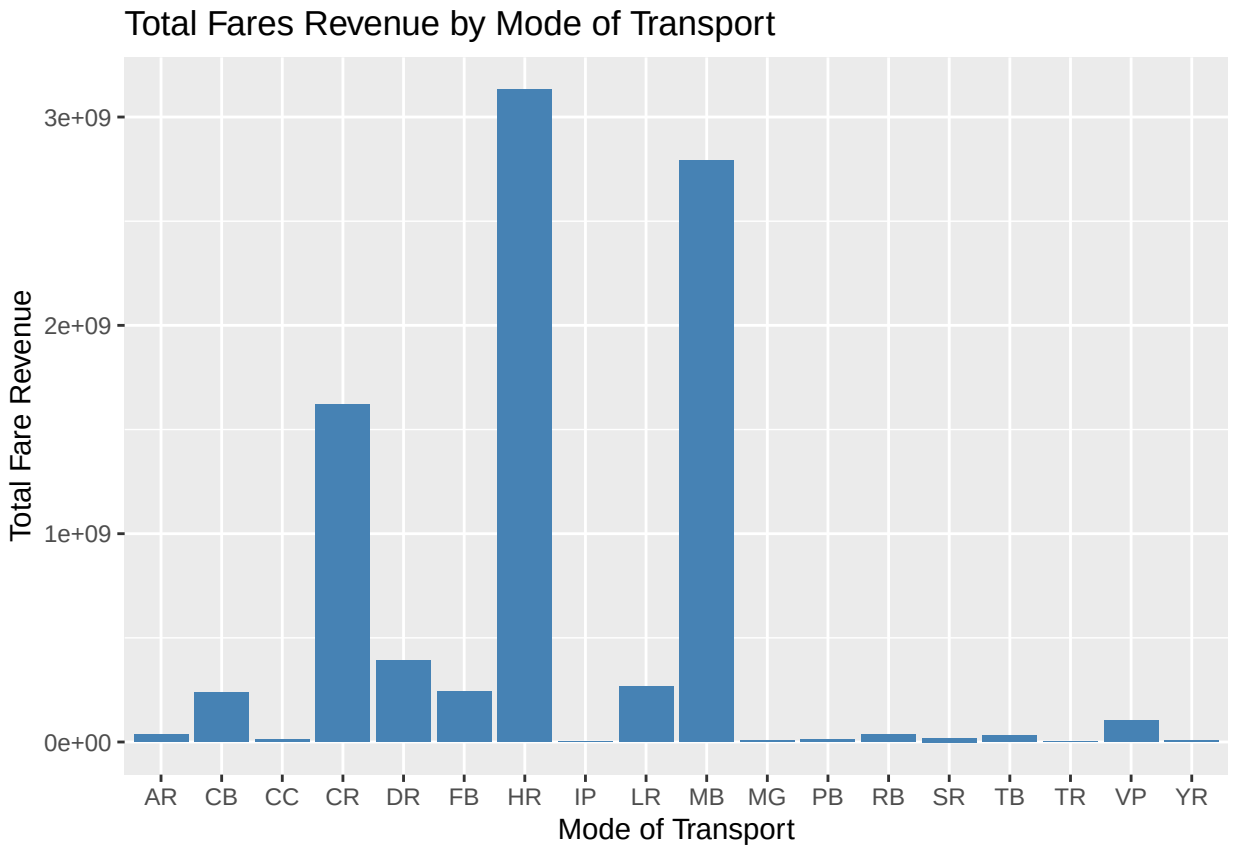
```
## [1] 0
```

Total fares earned by mode of transport

```
fares_by_mode <- df_clean_1 %>%
  group_by(mode) %>%
  summarise(fares_earned = sum(total_fares))

fares_by_mode
```

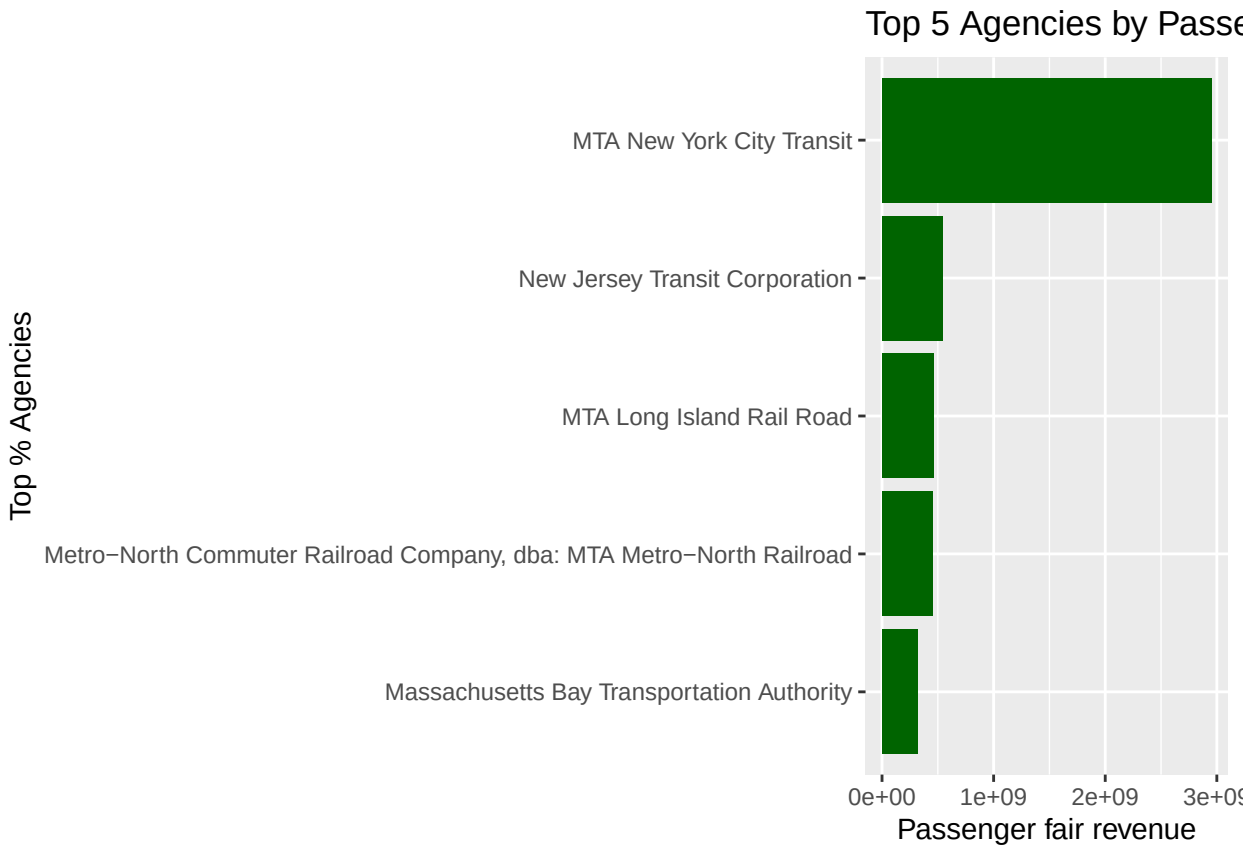
```
## # A tibble: 18 x 2
##    mode  fares_earned
##    <chr>        <dbl>
##  1 AR        33709263
##  2 CB       235707138
##  3 CC        10801075
##  4 CR      1618324270
##  5 DR       390731631
##  6 FB       240744994
##  7 HR      3129700829
##  8 IP         3407158
##  9 LR       268347252
## 10 MB      2792798130
## 11 MG         6526518
## 12 PB        11764061
## 13 RB        35473537
## 14 SR        19191862
## 15 TB        30025042
## 16 TR           80689
## 17 VP       103743926
## 18 YR         6747713
```

```
ggplot(fares_by_mode, aes(x = mode, y = fares_earned))+

  geom_bar(stat = 'Identity', fill='steelblue')+

  labs(title = "Total Fares Revenue by Mode of Transport", x = "Mode of Transport", y = "Total Fare Rev
```



Top 5 Agencies by Passenger Fares

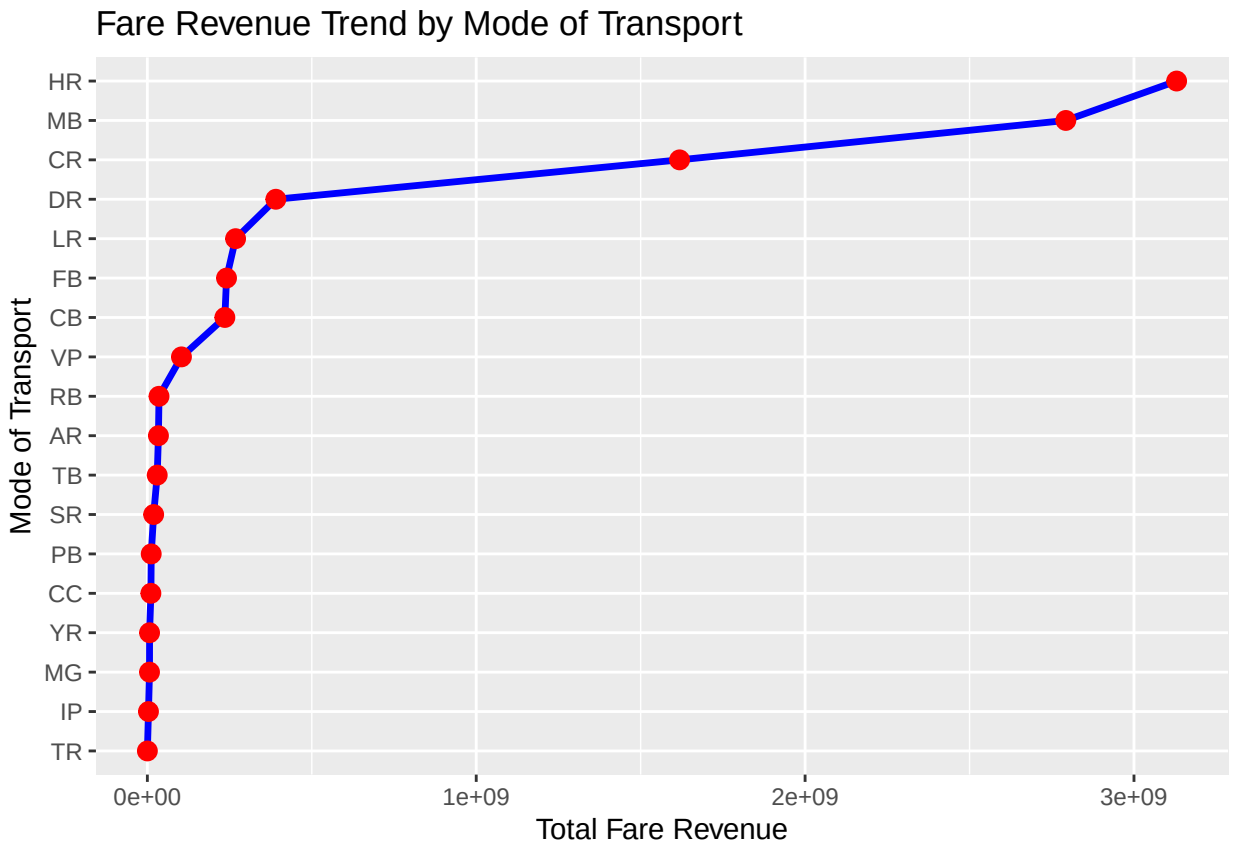
```
df_clean_1 %>%
  group_by(agency_name) %>%
  summarise(passenger_fares = sum(passenger_paid_fares)) %>%
  top_n(5, passenger_fares) %>%
  ggplot(aes(x = reorder(agency_name, passenger_fares), y = passenger_fares))+
  geom_bar(stat='identity', fill='darkgreen')+
  labs(title = "Top 5 Agencies by Passenger paid revenue", x = 'Top % Agencies', y = 'Passenger fair re
  coord_flip()
```



Fare_revenue trend across different modes

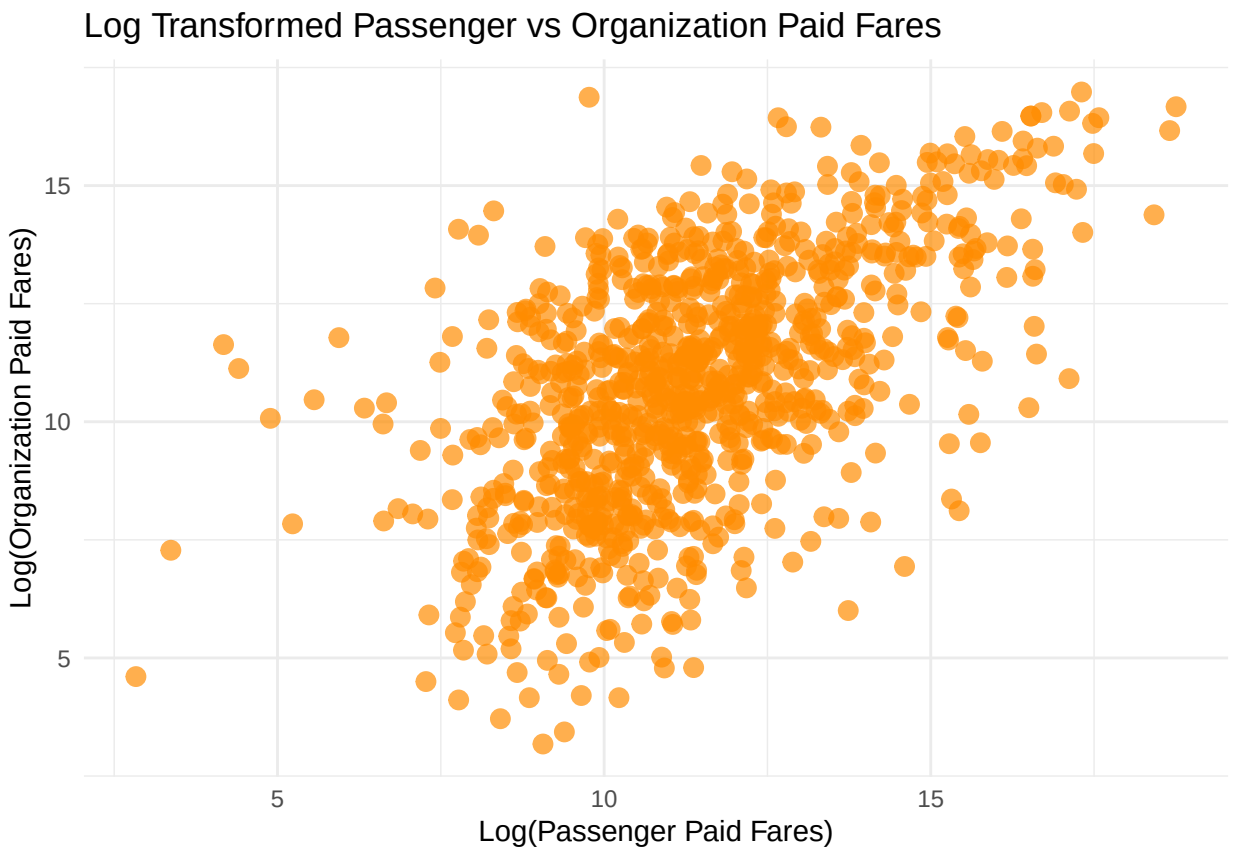
```
df_clean_1 %>%
  group_by(mode) %>%
  summarise(total_fares = sum(total_fares)) %>%
  arrange(desc(total_fares)) %>%
  ggplot(aes(x = reorder(mode, total_fares), y = total_fares, group = 1)) +
  geom_line(color = "blue", size = 1.2) +
  geom_point(color='red', size=3) +
  labs(title = "Fare Revenue Trend by Mode of Transport", x = "Mode of Transport", y = "Total Fare Reve
  coord_flip()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Passenger paid fare vs Organization Paid

```
df_clean_1 %>%
    filter(passenger_paid_fares > 0 & organization_paid_fares > 0) %>%
    mutate(
      log_passenger_paid_fares = log(passenger_paid_fares + 1),
      log_organization_paid_fares = log(organization_paid_fares + 1)
    ) %>%
    ggplot(aes(x = log_passenger_paid_fares, y = log_organization_paid_fares)) +
    geom_point(color = "darkorange", size = 3, alpha = 0.7) +
    labs(
      title = "Log Transformed Passenger vs Organization Paid Fares",
      x = "Log(Passenger Paid Fares)",
      y = "Log(Organization Paid Fares)"
    ) +
    theme_minimal()
```


Log Transformed Passenger vs Organization Paid Fares
Log(Organization Paid Fares)
15
10
5
5
10
15
Log(Passenger Paid Fares)