Assignment 3

Ashish Mhatre

04/02/2022

Problem Statement

Perform data analysis and visualization on Wine dataset, Farmer Market dataset and Airline delays dataset using data wrangling techniques.

Import Libraries

Importing all required modules

Task 1

Perform 5 subtask on Wine dataset.

Subtask 1

Find region in france having the highest average price.

```
## # A tibble: 1 x 2
##   region_1   Average_Price
##   <chr>              <dbl>
## 1 Montrachet          601.
```

Montrachet region has the highest average price of 601.1818

subtask 2

Find average price by designation

```
## # A tibble: 26,859 x 2
##    designation                                Average_Price
##    <chr>                                             <dbl>
##  1 Clos du Mesnil                                     1400
##  2 Roger Rose Vineyard                                1022.
##  3 Colheita White                                      980
##  4 El Perer                                            770
##  5 Les Quatre Journaux                                 740
##  6 Kiedrich GrÃ¤fenberg Trockenbeerenauslese           700.
```

```
##  7 Essencia                                          654
##  8 Hill of Grace                                     625
##  9 Figuero Tinus                                     599
## 10 La Cabotte                                        596
## # ... with 26,849 more rows
```

Following data set view shows the average price by designation

Subtask 3

Find the variety having the highest average price.

```
## # A tibble: 1 x 2
##   variety         Average_Price
##   <chr>                   <dbl>
## 1 Cabernet-Shiraz           150
```

Carbernet-Shiraz has the highest price of 150

Subtask 4

Display top 7 variety by count frequency.

```
## # A tibble: 7 x 2
##   variety                 Count
##   <chr>                   <int>
## 1 Chardonnay              14482
## 2 Pinot Noir              14291
## 3 Cabernet Sauvignon      12800
## 4 Red Blend               10062
## 5 Bordeaux-style Red Blend  7347
## 6 Sauvignon Blanc           6320
## 7 Syrah                     5825
```

Following dataset shows the top 7 variety by count frequency.

Subtask 5

Finding the number of wines which are 20 years old

```
## Count of 20 years old Wine is  84
```

There are 84 wines which are 20 year old.

Task 2

Generate wide table showing number of farmers market by state and month of year.

```
## # A tibble: 52 x 13
##    State January February March April   May  June  July August September October
##    <chr>   <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>     <dbl>   <dbl>
##  1 Alab~       9        0     1     7    24    21     1      1         0       2
##  2 Alas~       2        0     0     0     8     8     2      1         0       0
##  3 Ariz~      21        3     2     1    12     3     3      0         2      18
##  4 Arka~      10        0     3    11    27     6     0      0         0       0
##  5 Cali~     205        1    11    28    69    45     6      1         3       3
##  6 Colo~       5        0     0     2    18    53     8      0         0       0
##  7 Conn~       3        0     0     1     9    26     9      0         0       3
##  8 Dela~       1        0     0     2    11     8     0      0         0       1
##  9 Dist~       3        0     0     6    23    11     0      0         0       0
## 10 Flor~      84        1     3     5     7     4     0      2         3      15
## # ... with 42 more rows, and 2 more variables: November <dbl>, December <dbl>
```

We can see the number of farmers market for each month for a particular state in a single row.

Task 3

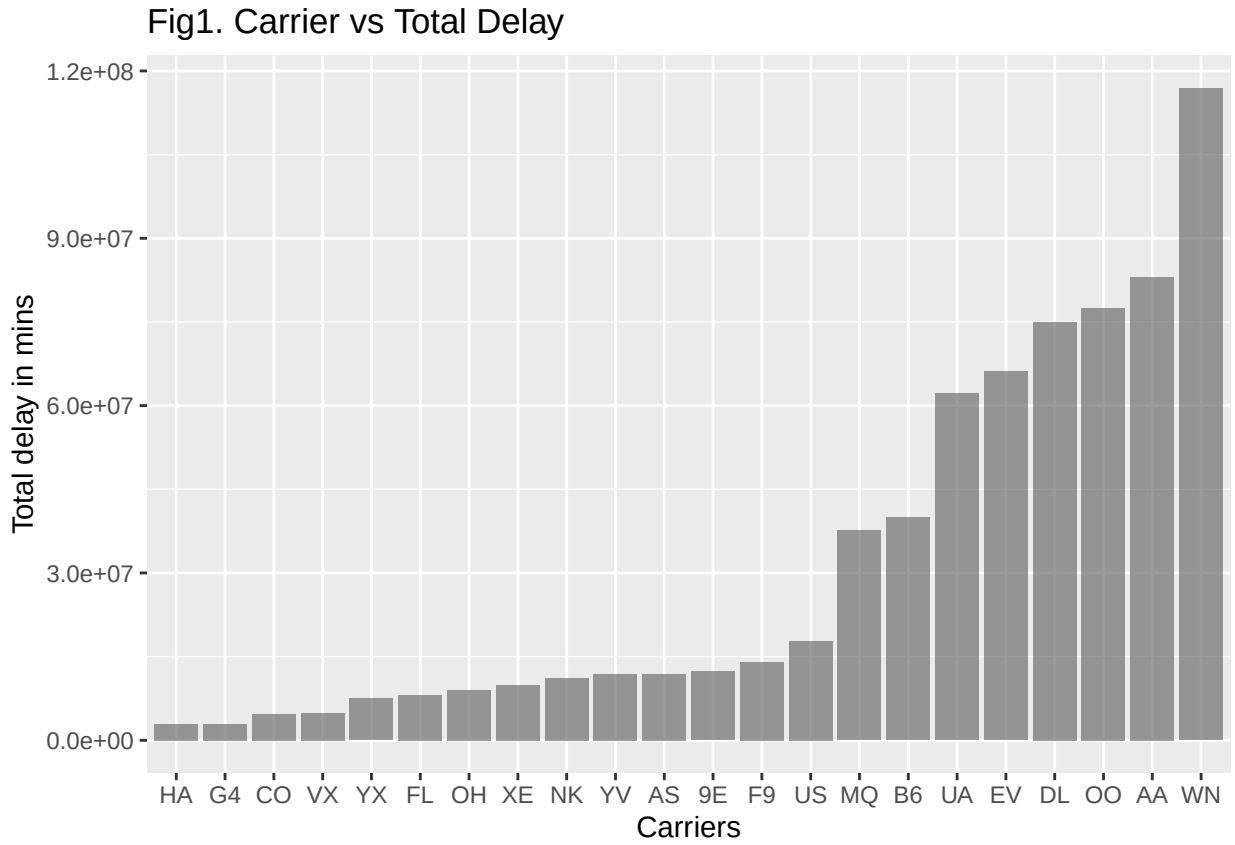
Showing the number of active farmer markets depending on the Updatetime column by month for each city in state of california

```
## # A tibble: 377 x 3
##    city             Update_month Active_Users
##    <chr>                   <dbl>        <int>
##  1 "San Francisco "            7            8
##  2 "Sacramento"                4            7
##  3 "San Francisco"             7            7
##  4 "San Jose"                  7            6
##  5 "San Diego"                12            5
##  6 "San Jose"                  6            5
##  7 "Anaheim"                   7            4
##  8 "Los Angeles"               6            4
##  9 "Oakland"                   5            4
## 10 "San Francisco"             6            4
## # ... with 367 more rows
```

Following dataset view provides us with the number of active farmers market for each city of california for each month.

Task 4

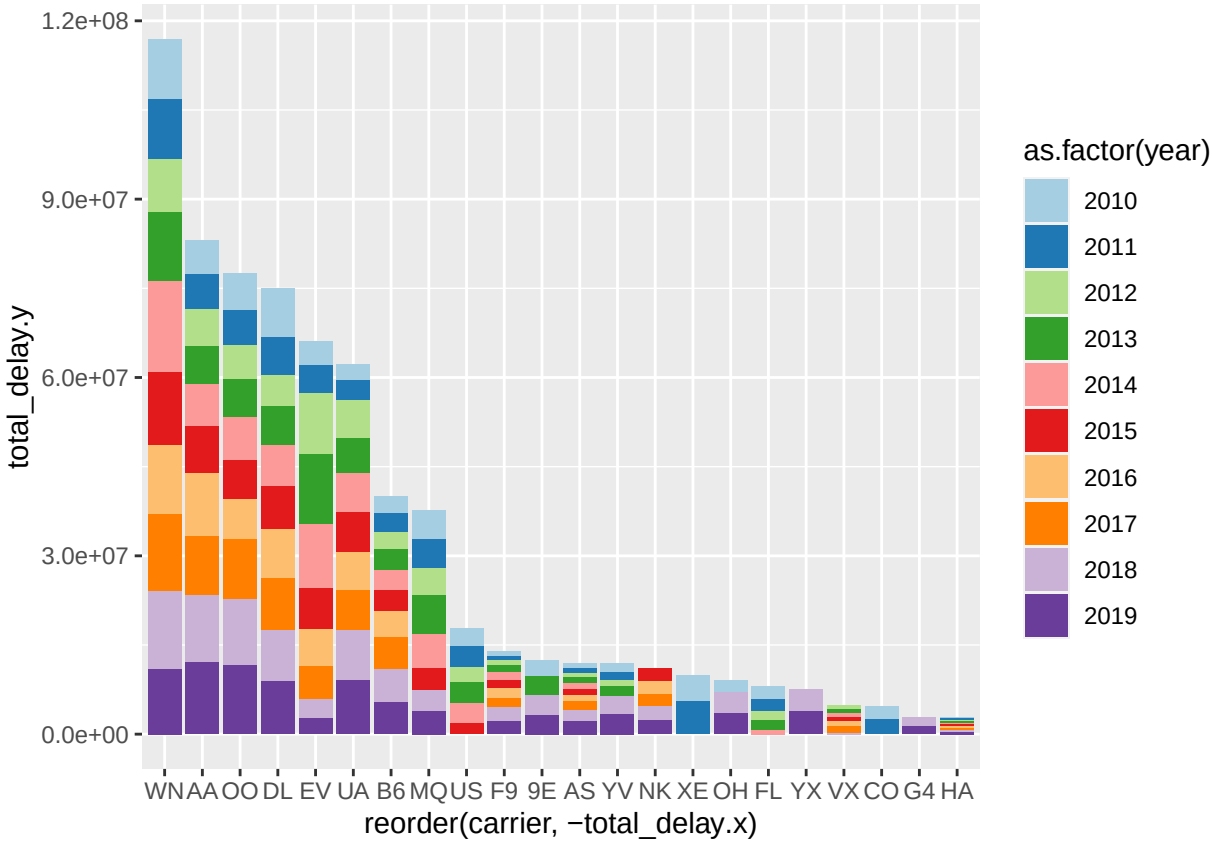
Plot a bar chart to display the total delay in mins by each carrier



We can see the distribution of total delays for each carrier where, x-axis represents the carrier name and y-axis represents the total delay in mins.

Task 4.1

Plot a stacked plot showing the total delay for each year stacked for a particular airline carrier.



The following graph shows carrier vs total delay by year, where x-axis represents the airline carrier and y-axis represent the time delay in mins. each year delay is represented with an different color as a stacked bar.

Conclusion

1. Performed following subtask :-

a. Montrachet region has the highest average price of 601.1818.

b. Displayed the average price by designation where in designation Clos du Mesnil has the highest average price 1400.00000.

c. Carbernet-Shiraz variety has the highest price of 150.

d. using group_by and summaries clause displayed data set view for the top 7 variety by count frequency.

e. using str_detect and reg-ex we found out that There are 84 wines which are 20 year old.

2. Generated view to see the number of farmers market for each month for a particular state in a single row using pivot_wider.

3. computed Number of active farmers market for each city of California for each month.where we found that San Francisco in the month of July had the most number of active markets that is 8.

4.We can see the distribution of total delays for each carrier where, x-axis represents the carrier name and y-axis represents the total delay in mins. where southwest airlines co. has the highest total delay compared to all airlines, also using (fill=year) argument in ggplot we created a stacked bar graph adding details for total delay by each year.