Project2_A

true
true

March 10, 2019

Introduction :-

As part of project 2 , where we have to take 3 different messy datasets and try to transform them using tifyr & dplyr and other libraries of R and prepare them for analysis so that we can infer something meaningfull from the same.

Problem Statment :-

This data set contains the avacado consumption of 30 cities throughout US, the data is scattered cross various excel(s) one each for one city. It has a lot of unwanted data which is messy and needs to be filtered out before to can use it to Analyse and infer something from the same.

Solution :-

We are using below libraries in our quest to resolve the above problem.:- *readxl*
dplyr
kableExtra
sqldf
stringr
ggplot2

1) Using the list.files method of readxl library , we are loading all the excel(s) in the folder with pattern ".xls" and using sapply function to read all files and bind_rows function add the file name to each row using name **"City"**

2) Using read.csv function we are loading another file region_city.csv which contains the city and to which region those city belongs information.

3) Then we exclude the unwanted columns from the first Data frame using select method of dplyr and apply mutate on that data frame to get the name of the city by removing the '.xls' and unwanted characters from the city name.

4) Using mutate we add a new column(Total Sales) which is the average price * total volume for that city.

5) Using the sqldf , we then combine the 2 data frames based on the city , now the new Data frame has everything including the region to which the city belongs.

6) Then we query to get the region & average Total Sales and group them by region.

7) Using KappleExtra library we display a few of the records of the final Data frame to be used for our analysis in Graph.

8) Using ggplot function of ggplot2 library we plot and show which region has the highest average total sales of Avacado .

```r
files <- list.files(pattern = "*.xls")
tbl <- sapply(files, read_excel, simplify=FALSE) %>%
bind_rows(.id = "City")
```

```
## readxl works best with a newer version of the tibble package.
## You currently have tibble v1.4.2.
## Falling back to column name repair from tibble <= v1.4.2.
## Message displays once per session.
```

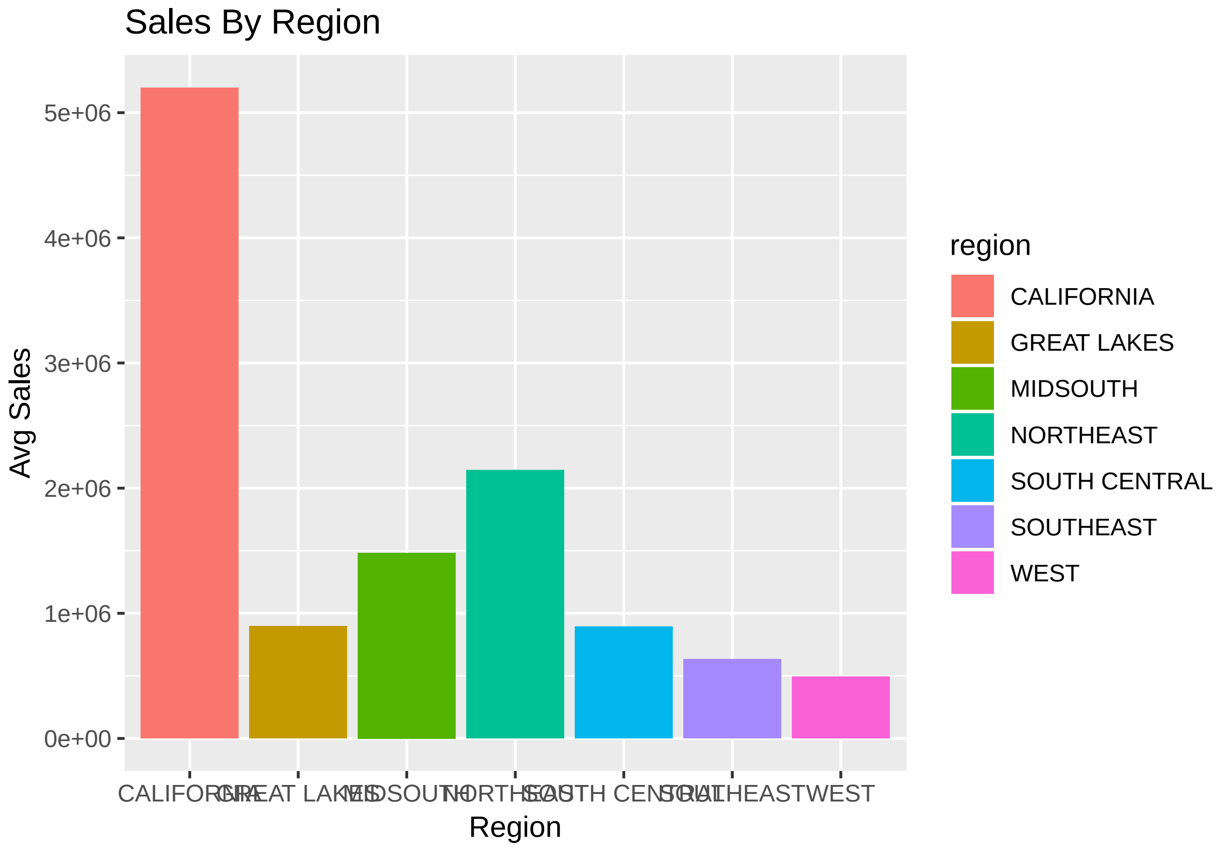
```r
kable(head(tbl)) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

City	Date	AveragePrice	Total
HAB_Retail_Volume_and_Price_2018_conventional_Atlanta.xls	2018-12-02	1.07	5
HAB_Retail_Volume_and_Price_2018_conventional_Atlanta.xls	2018-11-25	1.08	4
HAB_Retail_Volume_and_Price_2018_conventional_Atlanta.xls	2018-11-18	1.03	4
HAB_Retail_Volume_and_Price_2018_conventional_Atlanta.xls	2018-11-11	0.94	6
HAB_Retail_Volume_and_Price_2018_conventional_Atlanta.xls	2018-11-04	0.91	7
HAB_Retail_Volume_and_Price_2018_conventional_Atlanta.xls	2018-10-28	0.87	6

```r
region_table <- read.csv("region_city.csv")

tbl <- select(tbl , 1:4) %>%
          mutate(City = str_replace_all(City,"HAB_Retail_Volume_and_Price_2018_conventional_",""))  %>%
mutate(City = str_replace_all(City,".xls",""))
colnames(tbl) <- c("City","Date","AvgPrice","TotVol")

tbl <- mutate(tbl,Total_Sales = tbl$AvgPrice * tbl$TotVol)

combine_tbl <- sqldf("select a.City , a.total_sales , a.AvgPrice , b.region from tbl a left outer join 

graph_data <- sqldf("select region 'region' , round(avg(total_sales)) avg_sales  from combine_tbl
              group by region" )

kable(head(graph_data)) %>%
  kable_styling(bootstrap_options = c("striped","hover","condensed","responsive"),full_width   = F,posi
  row_spec(0, background ="gray")
```

region	avg_sales
CALIFORNIA	5199388
GREAT LAKES	896085
MIDSOUTH	1483060
NORTHEAST	2145480
SOUTH CENTRAL	892462
SOUTHEAST	634714

```r
graph_data %>%
  ggplot(aes(x=region, y=avg_sales, fill=region)) +
  scale_x_discrete(limits=graph_data$region) +
  geom_bar(stat="identity") +
  labs(title = "Sales By Region", x="Region", y="Avg Sales")
```



Summary :-

Using the graph we can clearly see that California region has the highest total sales of Avacado, followed by North East region and West region has the least consumption on avacado.