ETL and EDA in R

Loading the packages needed for EDA and ETL

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
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(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.0
## v purrr   1.0.2     v tibble  3.2.1
## v readr   2.1.4     v tidyr   1.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Loading the data

```
df <- read.csv('mod4peerrevdata.csv')
```

Transforming data types

Transforming Category to factor datatype

```
df$Category = as.factor(df$Category)
```

Transforming Date to mdy

```
df$Date = mdy(df$Date)
```

Displaying and interpreting the summaries for the Quantity and Price

```
summary(df$Quantity,df$Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    8.00   11.00   11.31   15.00   24.00       7
```

Counting NA values in each variable

```
summary(df)
```

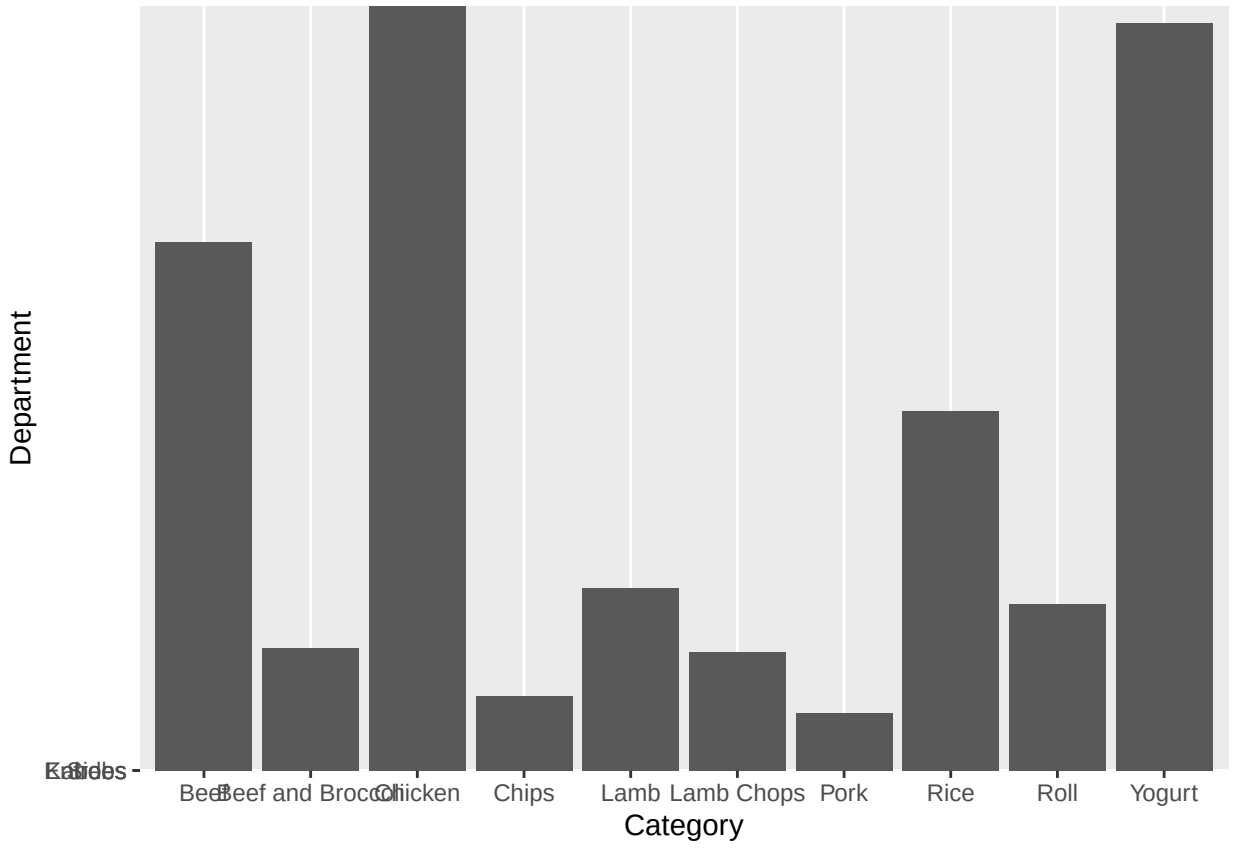
```
##       Date              Department                     Category
##  Min.   :2014-01-02   Length:34432       Chicken           :9046
##  1st Qu.:2014-09-11   Class :character   Beef              :6249
##  Median :2015-06-15   Mode  :character   Yogurt            :5898
##  Mean   :2015-07-05                      Beef and Broccoli:2885
##  3rd Qu.:2016-04-18                      Rice              :2835
##  Max.   :2017-04-03                      Lamb Chops        :2785
##                                          (Other)           :4734
##  CustomerCode           Price          Quantity
##  Length:34432       Min.   : 3.00   Min.   : 1.00
##  Class :character   1st Qu.:12.00   1st Qu.: 8.00
##  Mode  :character   Median :25.00   Median :11.00
##                     Mean   :22.81   Mean   :11.31
##                     3rd Qu.:33.00   3rd Qu.:15.00
##                     Max.   :50.00   Max.   :24.00
##                     NA's   :10      NA's   :7
```

```
df_new <- na.omit(df)
```

Bar chart for the Category column.

The bar chart should display the frequency of each category.

```
ggplot(df_new,aes(x=Category,y=Department)) +
  geom_col()
```



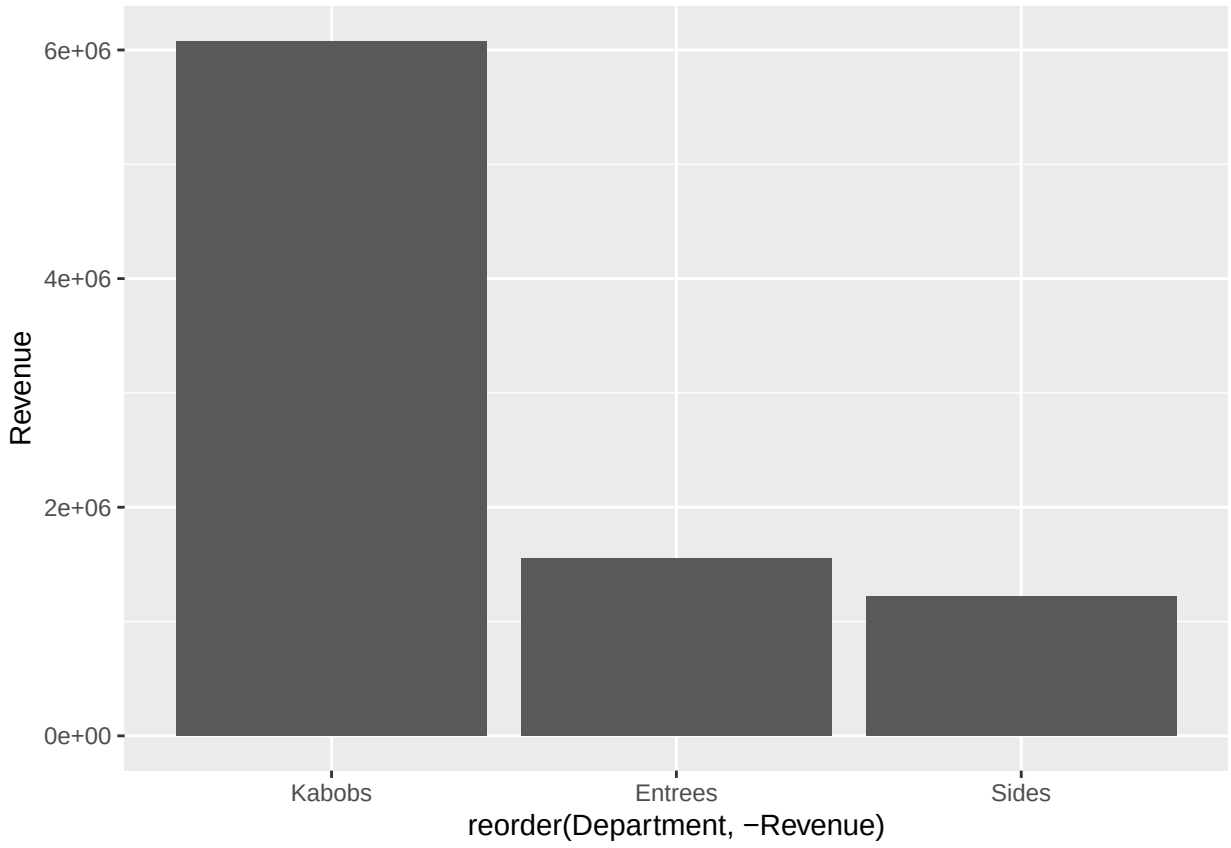
Bar chart of departments and their revenue

Create revenue

```
df_new$Revenue=(df_new$Price*df_new$Quantity)
```

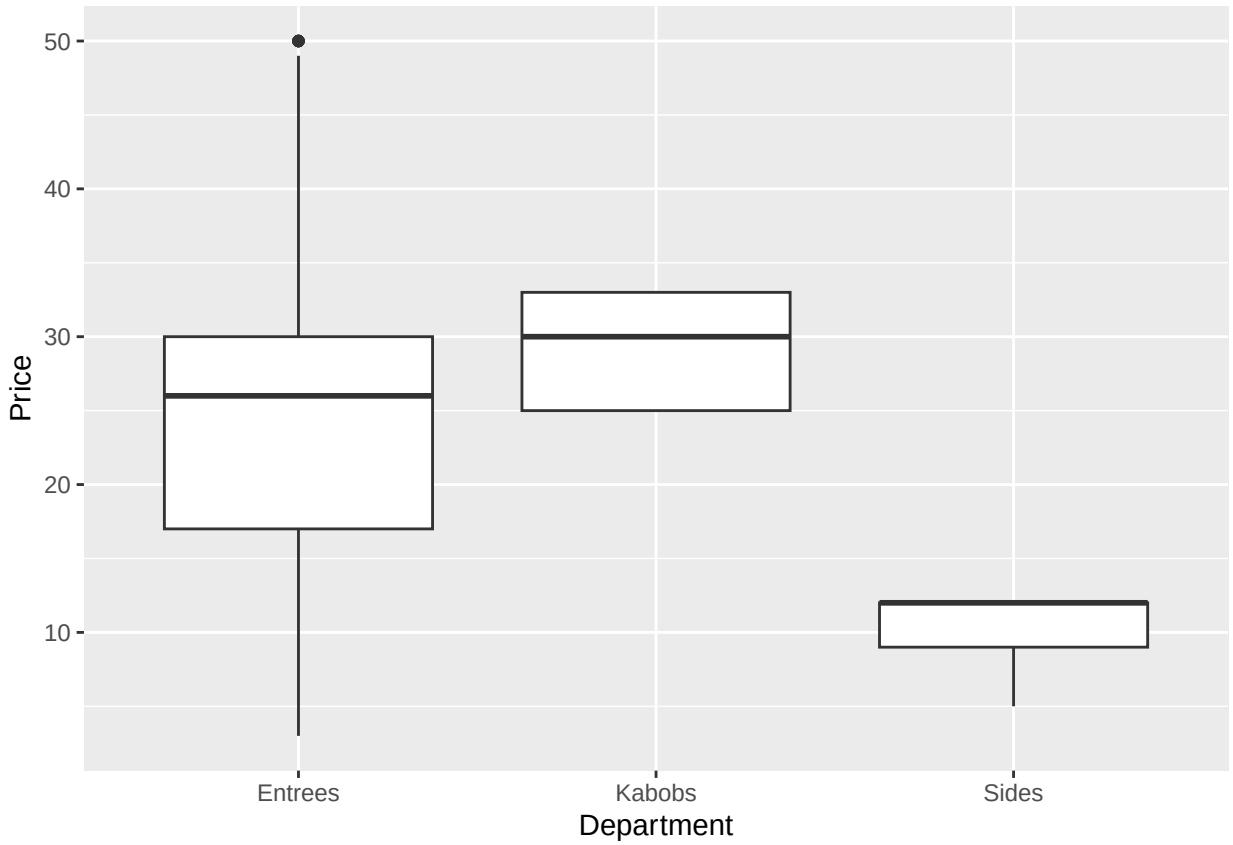
Create bar chart

```
ggplot(df_new,aes(x=reorder(Department,-Revenue),y=Revenue)) +
  geom_col()
```



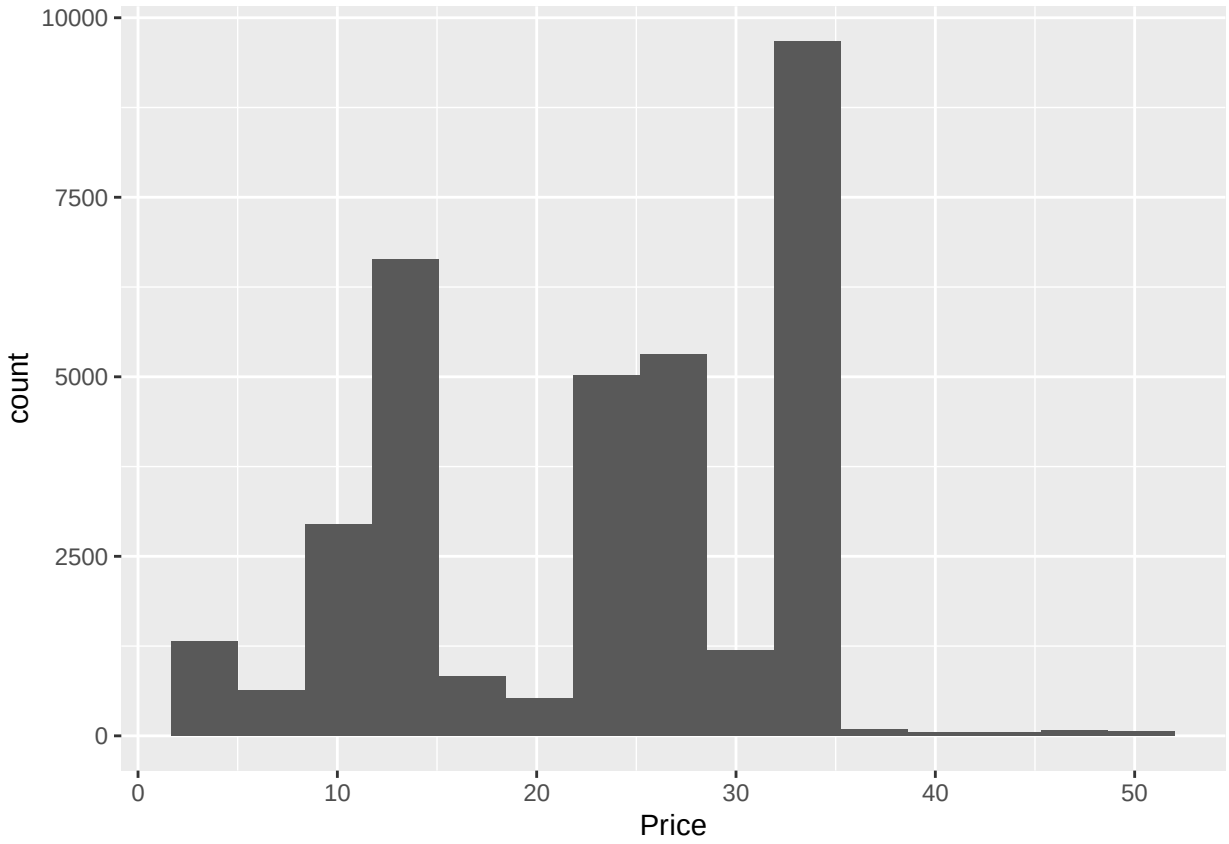
Box plot of the price column

```
ggplot(df_new,aes(x=Department,y=Price)) + geom_boxplot()
```



Histogram of the price column

```
ggplot(df_new, aes(Price)) +
 geom_histogram(bins = 15)
```

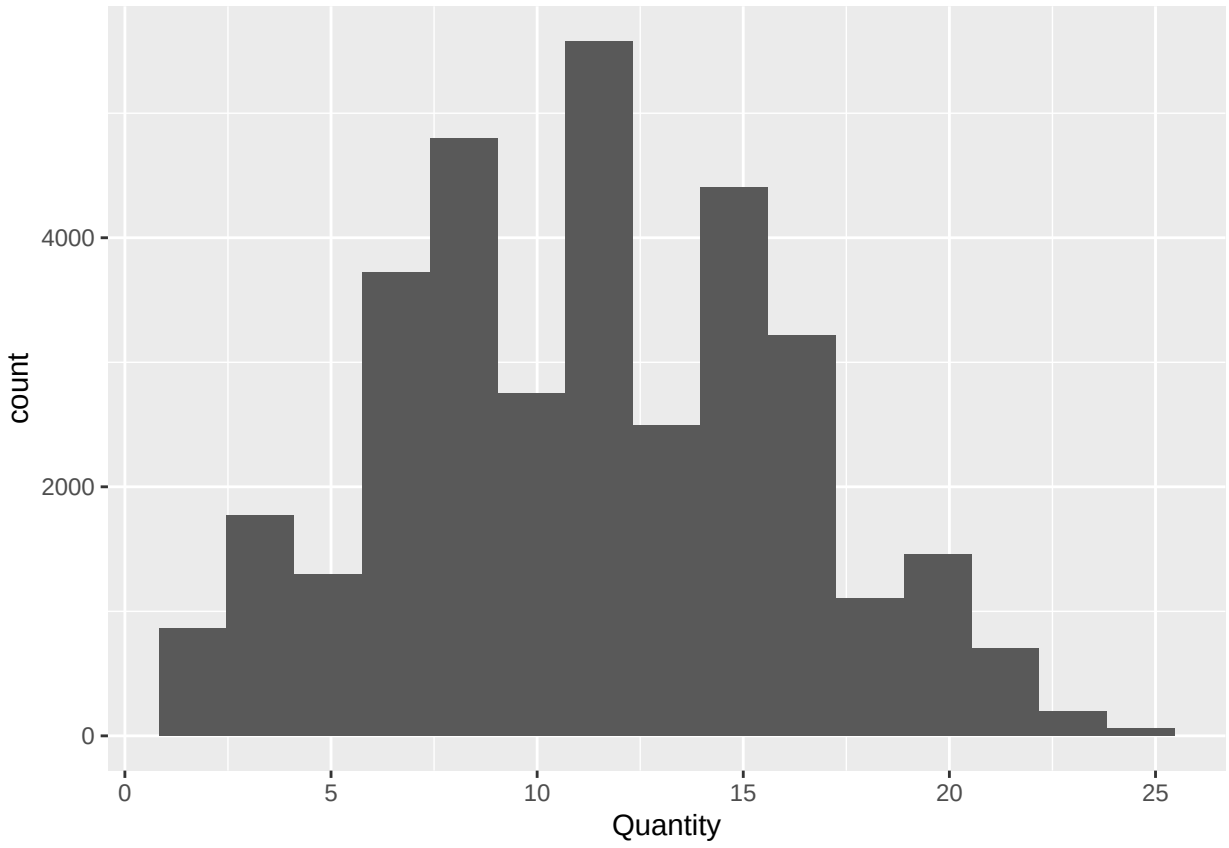


Box plot of the quantity column

```
ggplot(df_new,aes(x=Department,y=Quantity)) + geom_boxplot()
```

Histogram of the quantity column

```
ggplot(df_new, aes(Quantity)) +
 geom_histogram(bins = 15)
```



Comparison of PowerBI and Alteryx versus R for ETL and EDA.

Power BI, Alteryx, and R are powerful tools for analyzing revenue per department, each with its own strengths and weaknesses.

Power BI, a data visualization and business intelligence tool, excels in its user-friendly interface, making it easy for non-technical users to create interactive reports and dashboards. Its drag-and-drop functionality and seamless integration with other Microsoft products facilitate ease of use. However, Power BI's primary focus is on data visualization and reporting, limiting its advanced data transformation capabilities.

Alteryx is known for its data preparation and transformation capabilities. It allows users to clean and preprocess data effectively. While it may have a steeper learning curve compared to Power BI, it offers robust data analytics features, enabling in-depth revenue analysis. However, Alteryx might be more complex for beginners and comes at a higher cost.

R, on the other hand, offers unparalleled flexibility and statistical power. It's open-source, making it cost-effective, and provides advanced analytical capabilities. However, R often requires strong programming skills and data wrangling expertise, which can be a hurdle for non-technical users.

In terms of replicability, Alteryx and R offer better control and automation of data workflows. Power BI is more accessible for sharing results with non-technical stakeholders. Regarding scalability, Alteryx can handle large datasets efficiently, but R offers the most flexibility in terms of analysis. In summary, the choice among these tools depends on user expertise, project requirements, and available resources.