hw3-Deep-Bhalodia

Deep Bhalodia

2/5/2019

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Loading required packages

```
library(ggplot2)
library(dplyr)
library(rlang)
library(readr)
library(tidyr)
library(lubridate)
library(dbConnect)
library(dbplyr)
library(RMySQL)
library(DBI)
```

Part A

Problem 1

Find a dataset that is personally interesting to you

The dataset that is used for this problem can be founded on kaggle. The dataset belongs to e-commerce which contains information of actual transactions from UK retailer. It can be found here https://www.kaggle.com/carrie1/ecommerce-data/home

Import the dataset into R, put it into a tidy format, and print the first ten observations of the dataset.

```
data<-read_csv("data.csv")
```

```
glimpse(data)
```

```
## Observations: 541,909
## Variables: 8
## $ InvoiceNo   <chr> "536365", "536365", "536365", "536365", "536365", ...
## $ StockCode   <chr> "85123A", "71053", "84406B", "84029G", "84029E", "...
## $ Description <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHITE METAL...
## $ Quantity    <dbl> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2...
```

```
## $ InvoiceDate <chr> "12/1/2010 8:26", "12/1/2010 8:26", "12/1/2010 8:2...
## $ UnitPrice   <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1....
## $ CustomerID  <dbl> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 1...
## $ Country     <chr> "United Kingdom", "United Kingdom", "United Kingdo...
```

```
sum(is.na(data))
```

```
## [1] 136534
```

```
summary(data)
```

```
##   InvoiceNo          StockCode         Description
##  Length:541909      Length:541909      Length:541909
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Quantity         InvoiceDate          UnitPrice
##  Min.   :-80995.00   Length:541909      Min.   :-11062.06
##  1st Qu.:     1.00   Class :character   1st Qu.:     1.25
##  Median :     3.00   Mode  :character   Median :     2.08
##  Mean   :     9.55                      Mean   :     4.61
##  3rd Qu.:    10.00                      3rd Qu.:     4.13
##  Max.   : 80995.00                      Max.   : 38970.00
##
##    CustomerID       Country
##  Min.   :12346    Length:541909
##  1st Qu.:13953    Class :character
##  Median :15152    Mode  :character
##  Mean   :15288
##  3rd Qu.:16791
##  Max.   :18287
##  NA's   :135080
```

Comments

As seen from the summary statistics above there are NA values in column Customer ID. Also, the minimum value of quantity is negative which is not possible. We will deal with these in the following steps

In the next step we assign NA to columns Quantity and UnitPrice. As seen in the previous step there are negative values assigned to those column observations. Here we assign NA to all the negative values

```
data <- data %>%
  mutate(Quantity = replace(Quantity, Quantity<=0, NA),
         UnitPrice = replace(UnitPrice, UnitPrice<=0, NA))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

Here we drop all the rows containing NA values as discussed in the above steps

```
data <- data %>%
  drop_na()
```

Print first ten observations of the dataset

```
data[1:10,]
```

```
## # A tibble: 10 x 8
##    InvoiceNo StockCode Description Quantity InvoiceDate UnitPrice
##    <chr>     <chr>     <chr>          <dbl> <chr>           <dbl>
##  1 536365    85123A    WHITE HANG~        6 12/1/2010 ~      2.55
##  2 536365    71053     WHITE META~        6 12/1/2010 ~      3.39
##  3 536365    84406B    CREAM CUPI~        8 12/1/2010 ~      2.75
##  4 536365    84029G    KNITTED UN~        6 12/1/2010 ~      3.39
##  5 536365    84029E    RED WOOLLY~        6 12/1/2010 ~      3.39
##  6 536365    22752     SET 7 BABU~        2 12/1/2010 ~      7.65
##  7 536365    21730     GLASS STAR~        6 12/1/2010 ~      4.25
##  8 536366    22633     HAND WARME~        6 12/1/2010 ~      1.85
##  9 536366    22632     HAND WARME~        6 12/1/2010 ~      1.85
## 10 536367    84879     ASSORTED C~       32 12/1/2010 ~      1.69
## # ... with 2 more variables: CustomerID <dbl>, Country <chr>
```

InvoiceDate variable contains information about date and the time customer ordered something. Lets examine this variable in more detail

```
data$InvoiceDate[1:10]
```

```
##  [1] "12/1/2010 8:26" "12/1/2010 8:26" "12/1/2010 8:26" "12/1/2010 8:26"
##  [5] "12/1/2010 8:26" "12/1/2010 8:26" "12/1/2010 8:26" "12/1/2010 8:28"
##  [9] "12/1/2010 8:28" "12/1/2010 8:34"
```

The dates are in month/day/year hour:minute. At this point the variable InvoiceDate is a factor variable. This should be transformed into a datetime variable. To do so I have usee lubridate package.

```
# making two variables InvoiceDate and InvoiceTime
data <- separate(data, InvoiceDate, c("InvoiceDate", "InvoiceTime"), sep=" ", remove= TRUE)
data$InvoiceDate <- mdy(data$InvoiceDate) #make datetime object
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2019a.1.0/
## zoneinfo/America/New_York'
```

```
data$InvoiceTime <- hm(data$InvoiceTime) # make datetime objects
head(data)
```

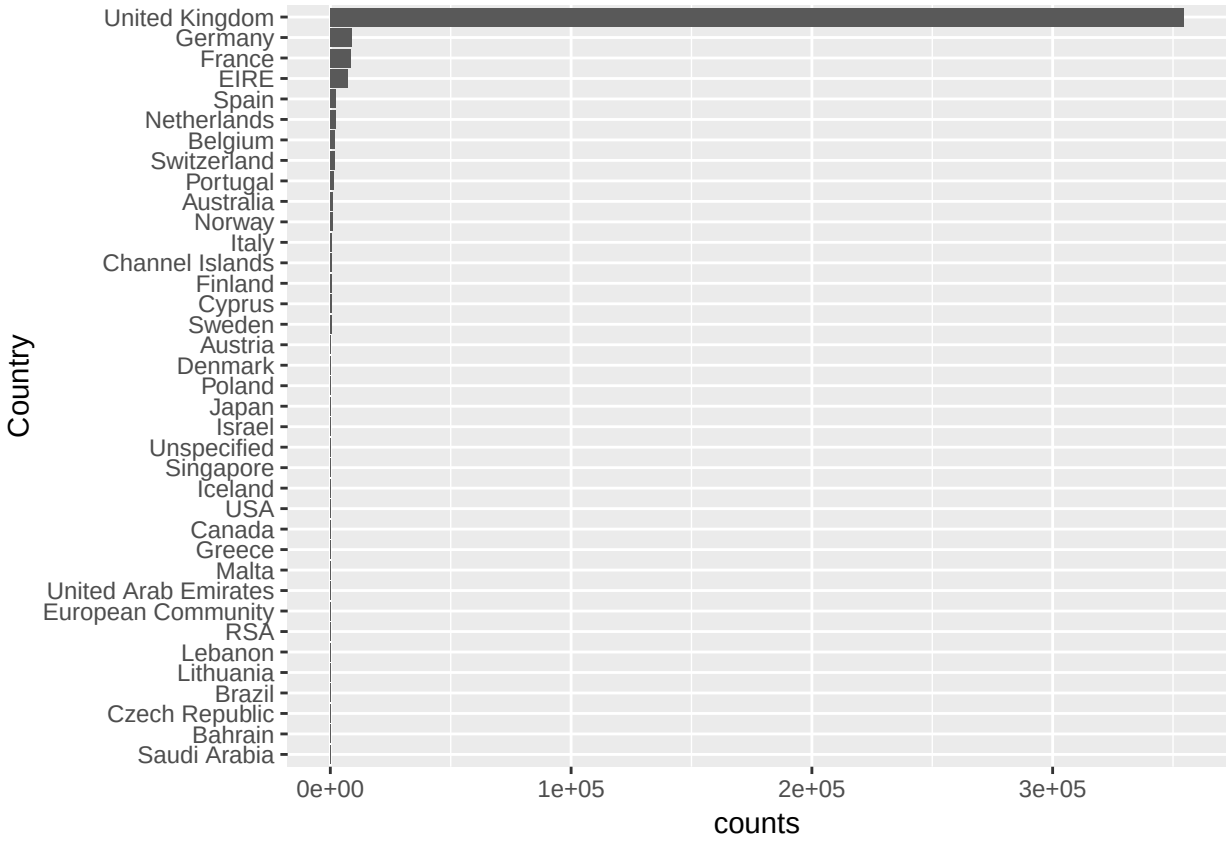
```
## # A tibble: 6 x 9
##   InvoiceNo StockCode Description Quantity InvoiceDate InvoiceTime
##   <chr>     <chr>     <chr>          <dbl> <date>      <S4: Perio>
## 1 536365    85123A    WHITE HANG~        6 2010-12-01  8H 26M 0S
```

```
## 2 536365    71053     WHITE META~        6 2010-12-01  8H 26M 0S
## 3 536365    84406B    CREAM CUPI~        8 2010-12-01  8H 26M 0S
## 4 536365    84029G    KNITTED UN~        6 2010-12-01  8H 26M 0S
## 5 536365    84029E    RED WOOLLY~        6 2010-12-01  8H 26M 0S
## 6 536365    22752     SET 7 BABU~        2 2010-12-01  8H 26M 0S
## # ... with 3 more variables: UnitPrice <dbl>, CustomerID <dbl>,
## #   Country <chr>
```

```
data$InvoiceYear <- year(data$InvoiceDate)
data$InvoiceMonth <- month(data$InvoiceDate,label=T)
data$InvoiceWeekday <- wday(data$InvoiceDate, label=T)
data$InvoiceHour <- hour(data$InvoiceTime)
```

Perform exploratory data analysis on the dataset, using the techniques learned in class. Calculate summary statistics that are of interest to you and create plots using ggplot2 that show your findings.

```
data %>%
  group_by(Country) %>%                                  # calculate the counts
  summarize(counts = n()) %>%
  arrange(counts) %>%                                    # sort by counts
  mutate(Country = factor(Country, Country)) %>%         # reset factor
  ggplot(aes(x=Country, y=counts)) +                     # plot
    geom_bar(stat="identity") +                          # plot histogram
    coord_flip()
```

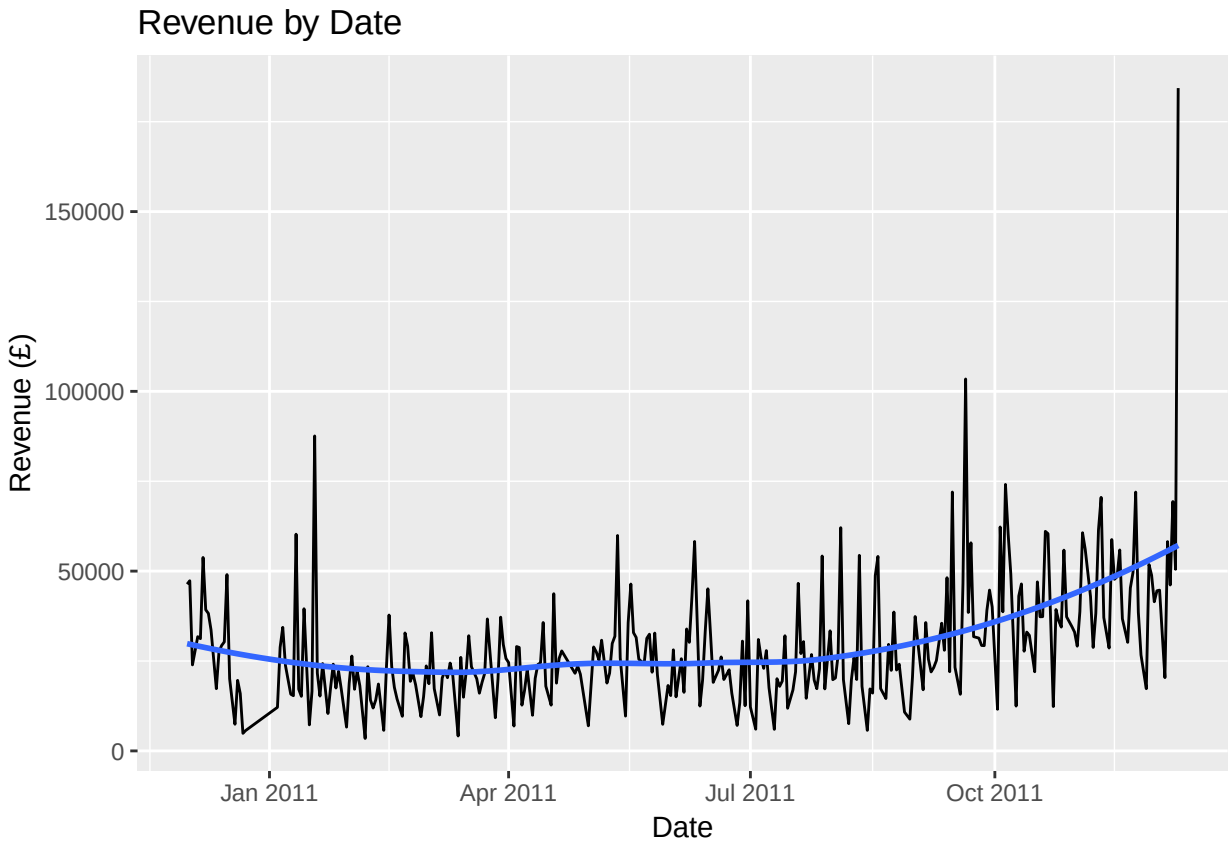


```
data <- data %>% mutate(lineTotal = Quantity * UnitPrice)
```

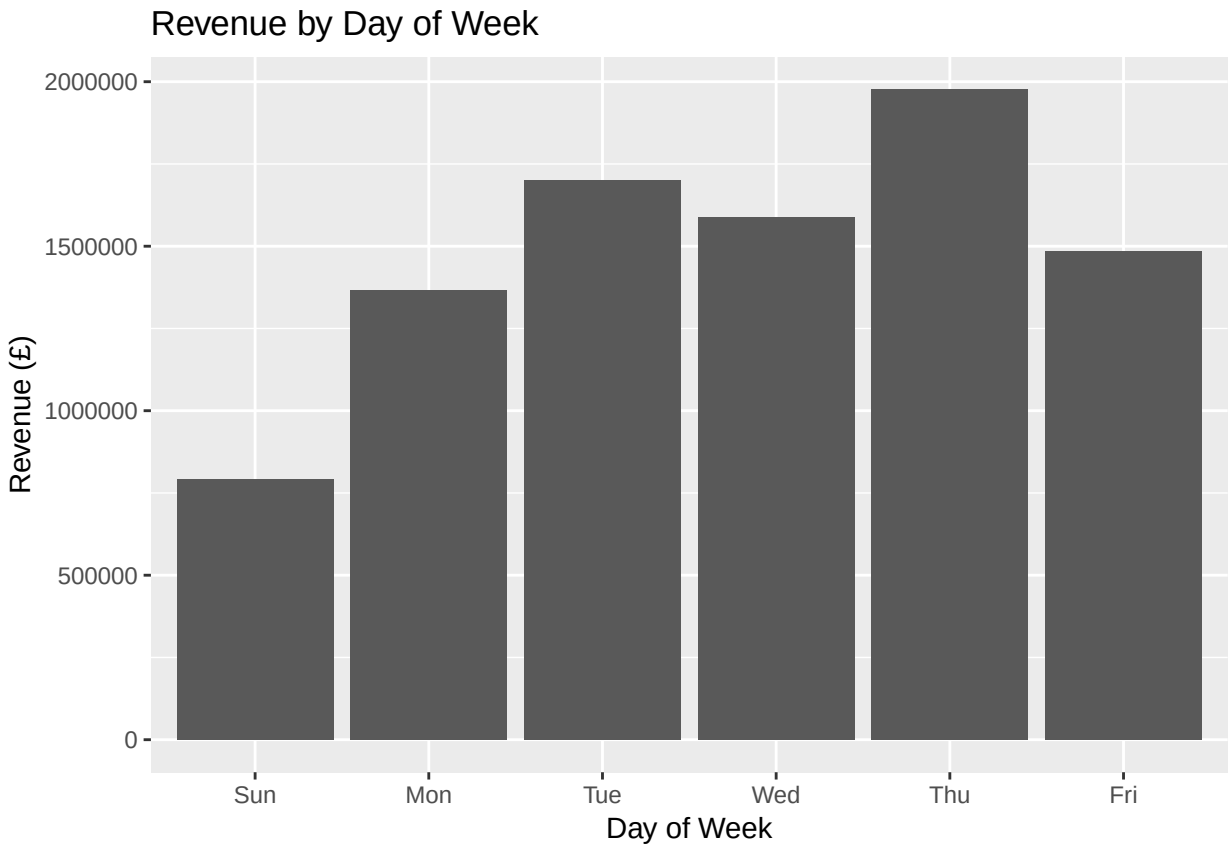
```
data$InvoiceYear<-as.factor(data$InvoiceYear)
data$InvoiceMonth<-as.factor(data$InvoiceMonth)
data$InvoiceWeekday<-as.factor(data$InvoiceWeekday)
data$InvoiceHour<-as.factor(data$InvoiceHour)
data$Country<-as.factor(data$Country)
```

```
options(repr.plot.width=8, repr.plot.height=3)
data %>%
  group_by(InvoiceDate) %>%
  summarise(Revenue = sum(lineTotal)) %>%
  ggplot(aes(x = InvoiceDate, y = Revenue)) + geom_line() + geom_smooth(method = 'auto', se = FALSE) + l
```

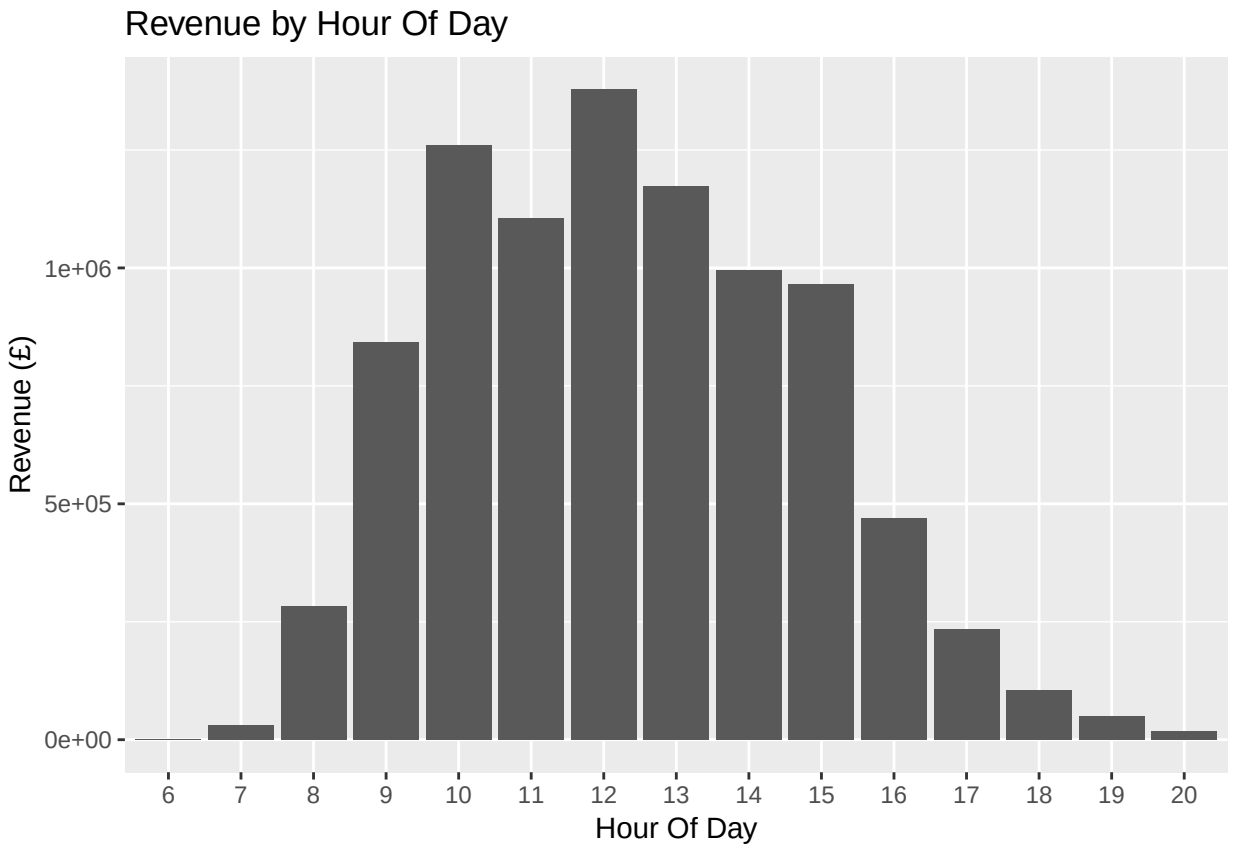
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



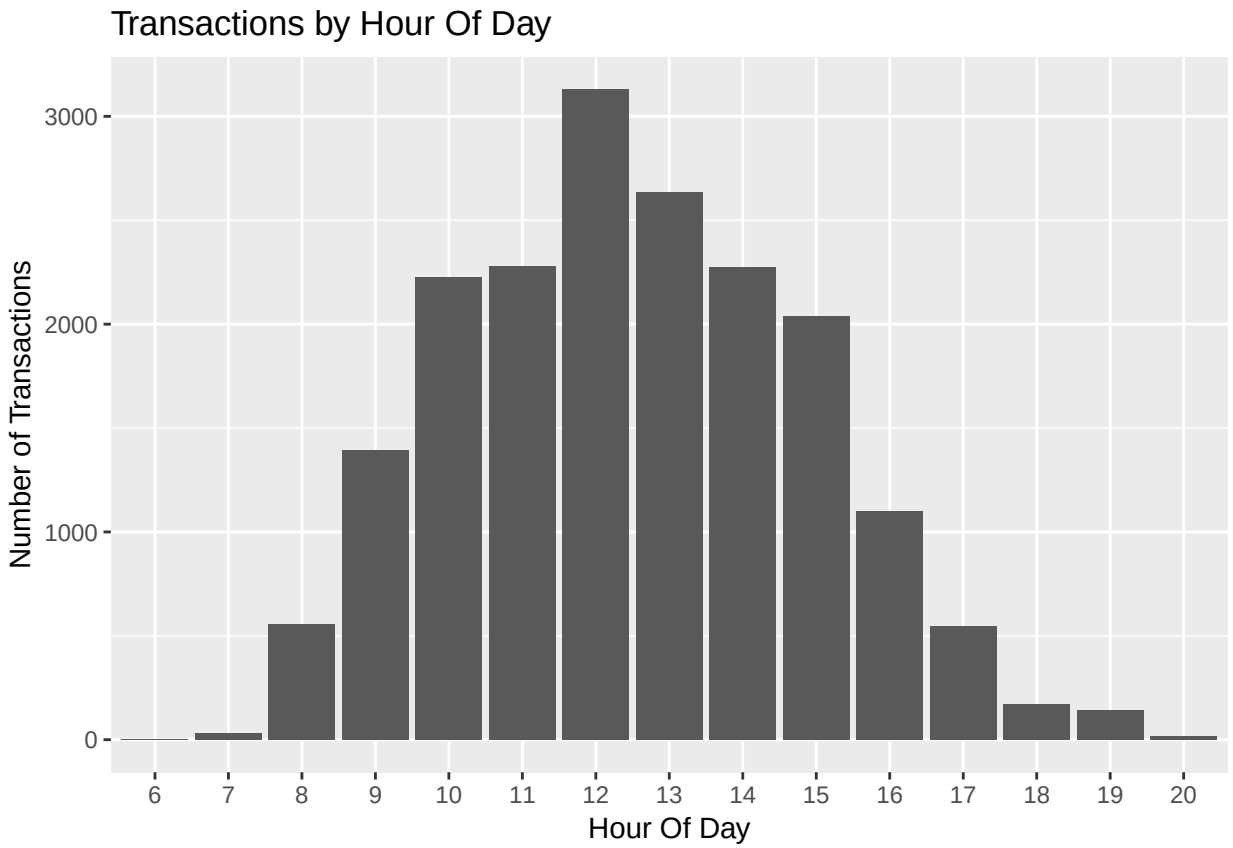

```
data %>%
  group_by(InvoiceWeekday) %>%
  summarise(revenue = sum(lineTotal)) %>%
  ggplot(aes(x = InvoiceWeekday, y = revenue)) + geom_col() + labs(x = 'Day of Week', y = 'Revenue (£)'
```




```
data %>%
  group_by(InvoiceHour) %>%
  summarise(revenue = sum(lineTotal)) %>%
  ggplot(aes(x = InvoiceHour, y = revenue)) + geom_col() + labs(x = 'Hour Of Day', y = 'Revenue (£)', t
```




```
data %>%
  group_by(InvoiceHour) %>%
  summarise(transactions = n_distinct(InvoiceNo)) %>%
  ggplot(aes(x = InvoiceHour, y = transactions)) + geom_col() + labs(x = 'Hour Of Day', y = 'Number of
```



Part B

Connecting with the MySql database

```
##Before reading data you should first create tables and load data in database. Then connect to the sam
con <- dbConnect(MySQL(), user="root", password="deep10", host="localhost", port=3306, dbname="DMDP")

dbListTables(con)
```

```
## [1] "authors" "general" "x"       "y"
```

Problem 3

Filter the data to include only the authors for whom a gender was predicted as 'male' or 'female' with a probability of 0.95 or greater, and then create a bar plot showing the total number of distinct male and female authors published each year. Comment on the visualization.

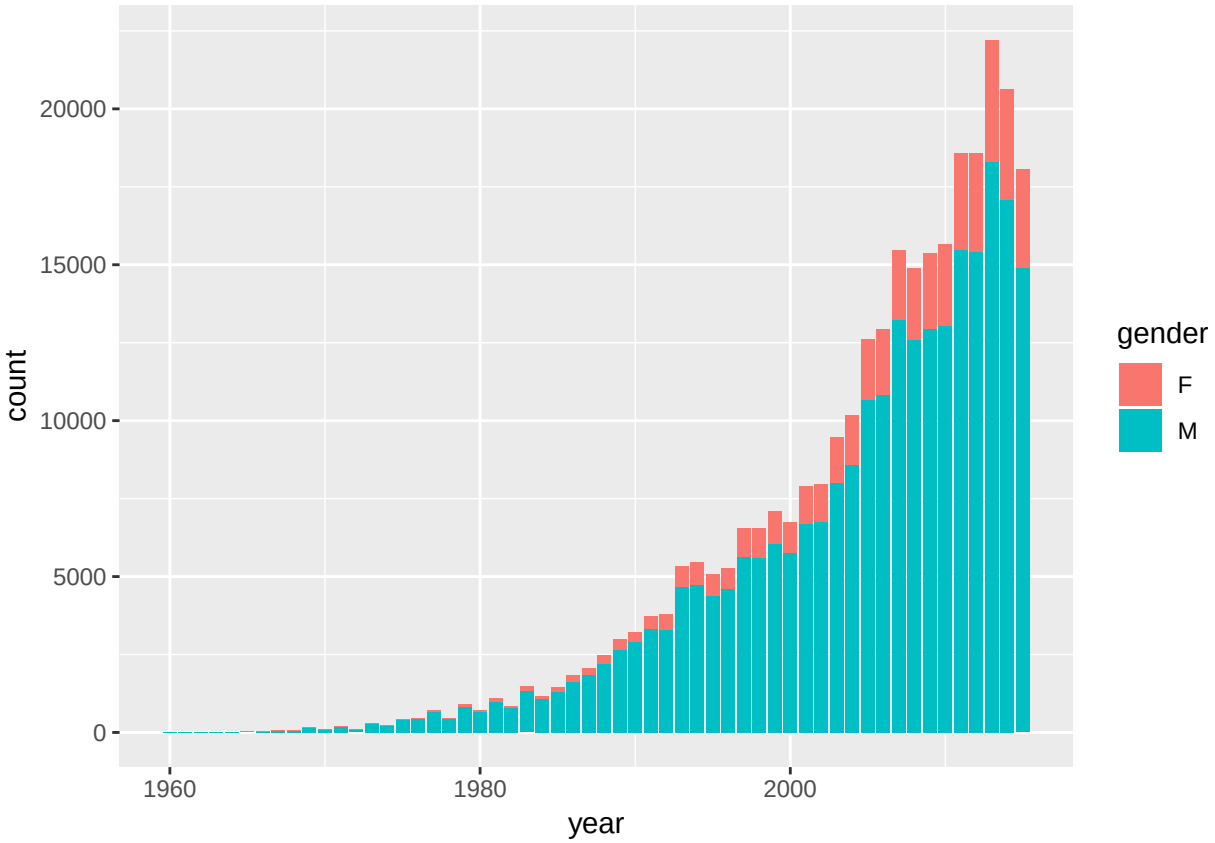
```
general <- tbl(con, "general")

authors <- tbl(con, "authors")
```

```
authors %>%
  left_join(general) %>%
  select(year, gender, prob, name, k) %>%
  collect() %>%
```

```
  filter(prob >= 0.95) %>%
  filter(gender %in% c('M', 'F')) %>%
  ggplot() +
  geom_bar(aes(x=year, fill = gender))
```

```
## Joining, by = "k"
```



Comments

Each row in the "authors" dataset corresponds to a single author on a single paper. Therefore, authors who have published more than one paper appear multiple times in the dataset. We use n_distinct() to count the number of distinct authors.

The total number of CS papers published each year is increasing over time. We also notice that the vast majority of authors publishing in computer science journals and proceedings each year are male.

Problem 4

Still including only the authors for whom a gender was predicted with a probability of 0.95 or greater, create a stacked bar plot showing the proportions of distinct male authors vs. distinct female authors published each year. (The stacked bars for each year will sum to one.) Comment on the visualization.

```
author_year <- authors %>%
  left_join(general) %>%
  select(k, year, name, gender, prob) %>%
  collect() %>%
```

```
  filter(gender %in% c('M', 'F')) %>%
  filter(prob >= 0.95) %>%
  group_by(year) %>%
  summarise(total = n_distinct(name))
```

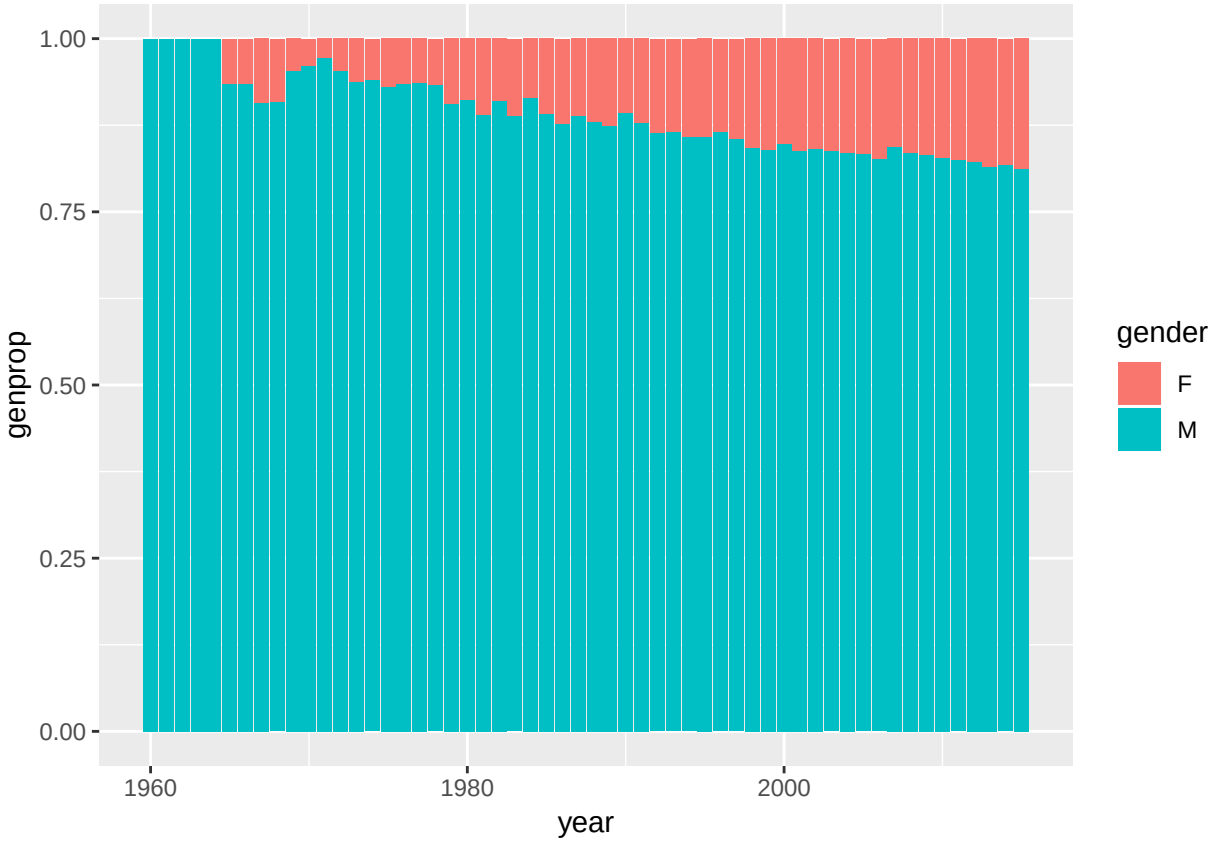
```
## Joining, by = "k"
```

```
general_author_year <- authors %>%
  left_join(general) %>%
  select(k, year, name, gender, prob) %>%
  collect() %>%
  filter(gender %in% c('M', 'F')) %>%
  filter(prob >= 0.95) %>%
  group_by(gender, year) %>%
  summarise(general_author_count = n_distinct(name))
```

```
## Joining, by = "k"
```

```
left_join(general_author_year, author_year) %>%
  select(year, total, gender, general_author_count) %>%
  mutate(genprop = general_author_count / total) %>%
  ggplot() +
  geom_col(aes(x=year, y=genprop, fill=gender))
```

```
## Joining, by = "year"
```



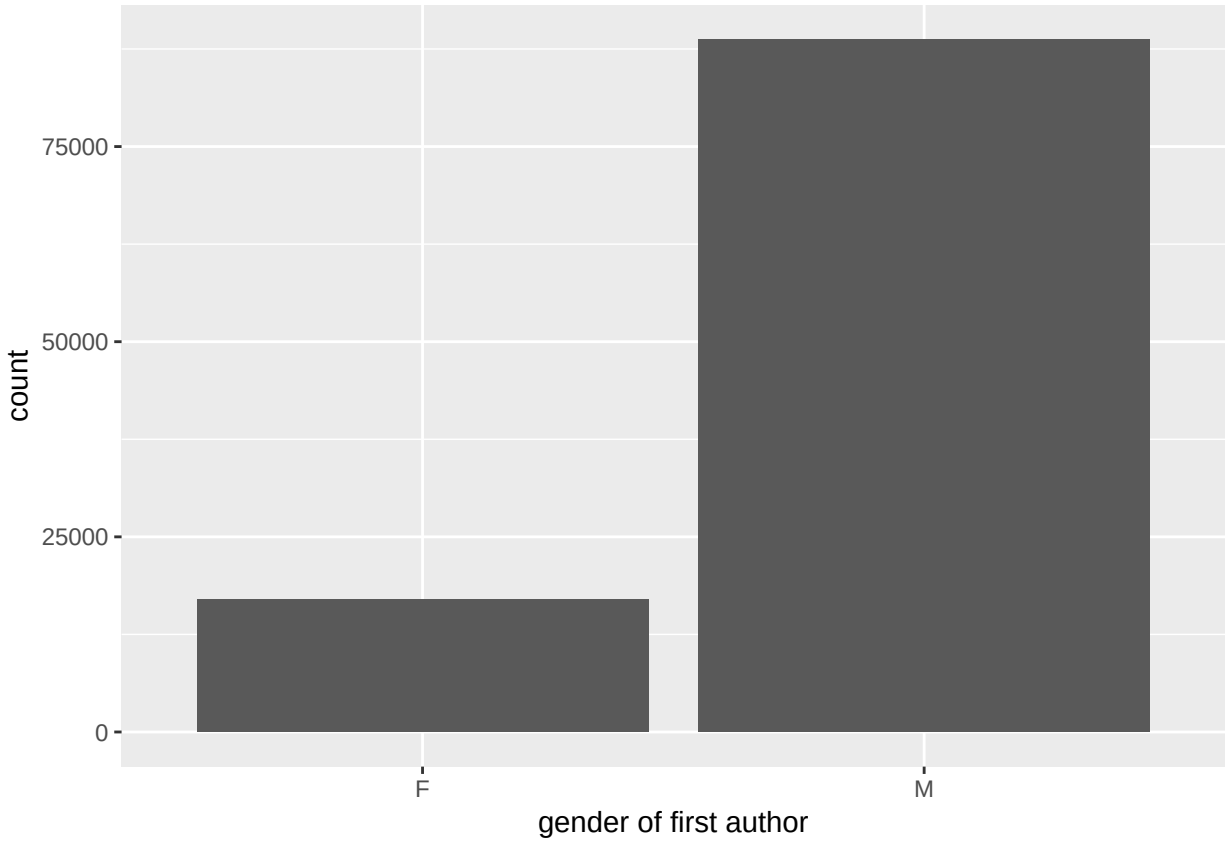
Comments

Because authors appear multiple times in the dataset, we calculate the year-by-year counts separetely, and then join the summaries with left_join() to plot the proportions. We see that there is a general trend of the proportion of women authors increasing over the years, but there is still a long way to go.

Problem 5

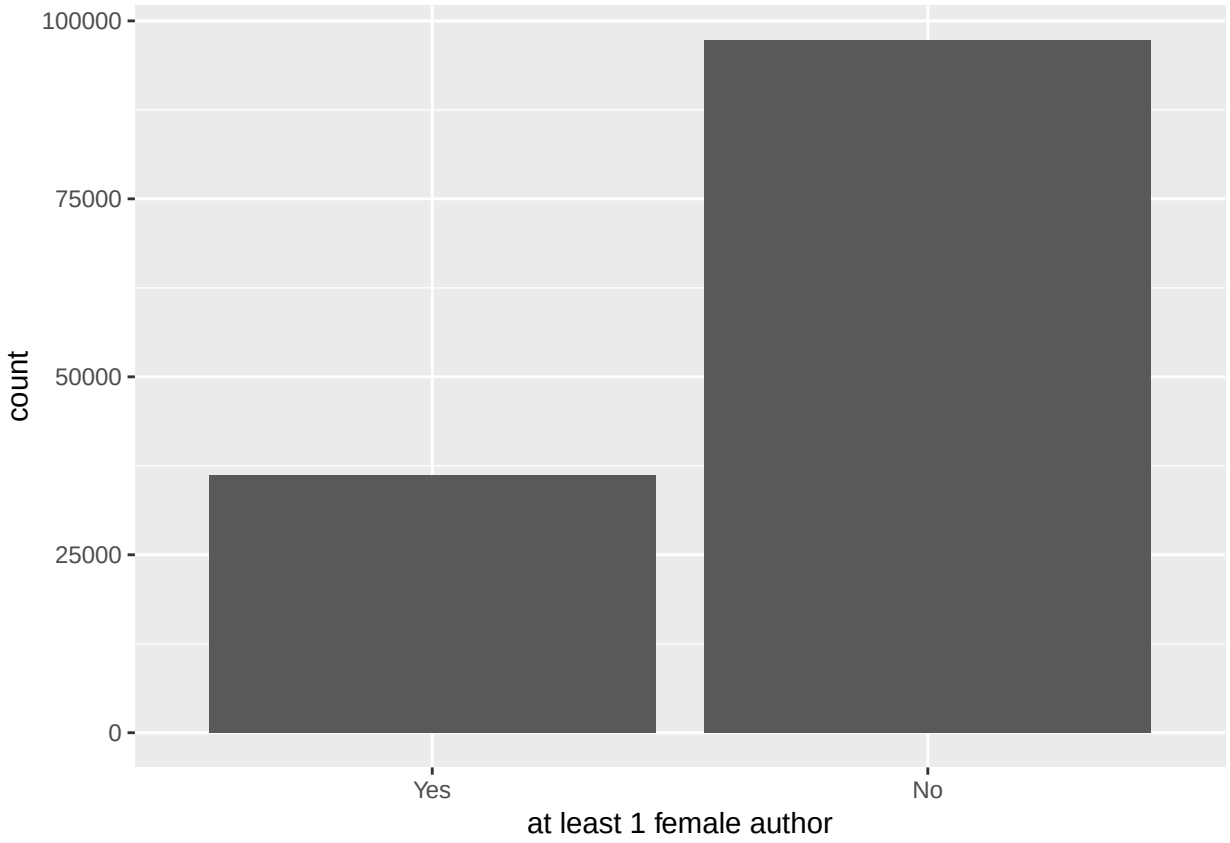
Still including only the authors for whom a gender was predicted with a probability of 0.95 or greater, create a bar plot showing the count of papers published with (1) male first authors and (2) female first authors. Then create a bar plot showing the count of papers published with (1) no female authors and (2) at least 1 female author. Comment on any similarities and differences between the two bar plots.

```
#Part-1
authors %>%filter(gender == 'M' | gender == 'F', prob >= 0.90) %>%
  filter(pos == 0) %>%
  collect() %>%
  group_by(gender) %>%
  ggplot() +
  geom_bar(aes(x=gender)) +
  labs(x="gender of first author")
```




```
#Part-2
authors %>%filter(gender == 'M' | gender == 'F', prob >= 0.95) %>%
  collect() %>%
```

```
  group_by(k) %>%
  summarise(AnyF = "F" %in% gender) %>%
  mutate(AnyF = factor(AnyF, levels = c("TRUE", "FALSE"),
  labels = c("Yes", "No"))) %>%
  ggplot() +
  geom_bar(aes(x=AnyF)) +
  labs(x="at least 1 female author")
```



Comments

The first plot shows there are far fewer CS papers published with female first authors than male first authors.

The second plots shows that – although there are still fewer papers published with any female authors than without female authors – the difference is less than before.

This suggests there are more women publishing CS papers as co-authors than first authorships alone would suggest.