Assignment 8

Instructions

- Please produce your assignment as a pdf (knit to pdf). See instructions in announcements if you have not downloaded a LaTeX distributor. If you are still having issues, you may knit to HTML and use your browser to produce a pdf file, as it is detailed in the M 01 09 video of module 1.
- You will have to produce this (and future) qmd file.
- You do not need to copy the statements.
- Submit your qmd file (it will not be graded but we want it for reference purposes).
- Show all the code (use echo = TRUE as option in R chunks) as well as the results.
- Each exercise is worth 25 points.
- For problems 1-3, please use the database complaintsdata.db, which is given with this pdf in assignment 8.
- Refer to the Class SQL notes, but use something similar to the code below to get started and connect to the database. Note that this file is db and not in sql lite.

Exercise 1 SQL Basics: [25]

Define the variable dcon as follows (using the consumercomplaint.db database).

```
library(knitr)
library(RSQLite)
dcon <- dbConnect(SQLite(), dbname = "consumercomplaint.db")
```

1. List the sql tables for dcon. Consumer_complaints table must show, the other tables do not need to show. (Note: Use dbListTables). (4 points)

```
## Begin Solution

## End Solution
```

2. List the fields in the sql table consumer_complaints. (Note: Use dbListFields). (6 points)

```
## Begin Solution

## End Solution
```

3. Create a dataframe called mydf by selecting product, sub_product, state, and ordering by Year and then Month. Use the head function to show the first few rows of mydf. (Note: You are using sql to select the field names).(15 points)

```
## Begin Solution

## End Solution
```

Exercise 2 SQL Filtering: [25]

1. Use the sql database and the consumer_complaints sql table,

    a. Select date_received, product, sub_product and filter where Year is greater than 2013. Order by date_received, product, and sub_product. Use head function to show first few rows. (5 points)

```
## Begin Solution

## End Solution
```

b. Select company, product and filter where state is TX and order by company and product. Use head function to show first few rows. (5 points)

```
## Begin Solution

## End Solution
```

c. Select product, submitted_via, and filter where submitted_via is Email and order by product. Use head function to show first few rows. (5 points)

```
## Begin Solution

## End Solution
```

2. Perform the following.

a. Select product and month and filter where month is between 'A' and 'J'. Order by product and month. Use head function to show first few rows. Note: months should be ordered alphabetically for this assignment, not chronologically. (5 points)

```
## Begin Solution

## End Solution
```

b. Select company and product and filter where state is like 'C%'. Order by company and product. Use head function to show first few rows. (5 points)

```
## Begin Solution

## End Solution
```

Exercise 3 Aggregrate Functions: [25]

a. Use the sql database and the sql table consumer_complaints and count the number of observations for submitted_via (4 points). Use GROUP BY submitted_via.

```
## Begin Solution

## End Solution
```

b. Use the sql database and the sql table consumer_complaints and count the number of observations for product and submitted_via. (Note: Group by product and submitted_via) (6 points).

```
## Begin Solution

## End Solution
```

c. Use the sql database and the sql table consumer_complaints and count the number of observations for product and submitted_via based on the conditions where Year is greater than 2013, submitted_via is Web, and the state is TX. (Note: Use group by for product and submitted_via) (15 points)

```
## Begin Solution

## End Solution
```

Exercise 4: [25]

Update your group project by creating a database with your main source and at least one other secondary source. Make some first queries to reduce the size of the data.frame (operating under the assumption that you are working with a huge dataset that would not fit in physical memory, even if this is not the case) so that you would be able to continue transforming with dplyr. Dplyr is not necessary on this homework.

- This will be a group exercise, to be uploaded by only one member of each group
- If you are not the group member that uploaded the solution, please specify which member of your group uploaded it
- Only one submission for the group.