Instructions

Level 1 Parts

Level 2 parts:

P2.5: Dataset name `data` in Problem 2.

Problem 4: Baby name visualization

Include the person you coop with:

Appendix

STA 032 Homework 3

CHANGE YOUR NAME HERE

DUE: May 3 2023, 12PM

Instructions

- Upload a PDF file, named with your UC Davis email ID and homework number (e.g., `xjw18_hw3.pdf`), to Gradescope (accessible through Canvas). You will give the commands to answer each question in its own code block, which will also produce output that will be automatically embedded in the output file. When asked, answer must be supported by written statements as well as any code used.
- All code used to produce your results must be shown in your PDF file (e.g., do not use `echo = FALSE` or `include = FALSE` as options anywhere). `Rmd` files do not need to be submitted, but may be requested by the TA and must be available when the assignment is submitted.
- Students may choose to collaborate with each other on the homework, but must clearly indicate with whom they collaborated. Every student must upload their own submission.
- Start to work on it as early as possible. Finishing this homework can help prepare midterm 1.
- When you want to show your result as a vector that is too long, slice the first 10 objects. When you want to show your result as a data frame, use `head()` on it. Failure to do so may lead to point deduction.
- Directly knit the Rmd file will give you an html file. Open that file in your browser and then you can print it into a PDF file.

Some Level 2 problems will show in the level 1 parts because they share the same background, but please type your solution in Level 2 parts!!

Level 1 Parts

If you have not installed these packages, please install them by uncommenting the code. You only need to install the package once.

```
#install.packages("ggplot2")
#install.packages("tidyverse")
#install.packages("ggthemes")
#install.packages("ggrepel")
#install.packages("ggwordcloud")
```

Loading the required pacakges.

```
library(tidyverse)
library(ggplot2)
library(ggthemes)
library(ggwordcloud)
```

Problem 1: Conceptual problems

State True or False for each of the statements. Then provide your reason.

1. For a sample dataset, it is not possible for the first quartile to equal the median.
2. If the sample correlation for variable X and Y is zero, then the variable X and Y does not have any relationship.

Instructions

Level 1 Parts

Level 2 parts:

P2.5: Dataset name `data` in Problem 2.

Problem 4: Baby name visualization

Include the person you coop with:

Appendix

3. The standard deviation must always be larger than the mean.
4. Outliers have a strong influence on the standard deviation of a dataset.
5. The salaries of all persons employed by a large university is left skewed.
6. The amounts of the time spent by students on a difficult examination is left skewed.
7. We use R to calculate the mean and variance of the student heights in the dataset `heights`. Mean is `mean(heights$height)=68.32` and variance is `var(heights$height)=16.64`. So the population mean is 68.32 and population variance is 16.64.

Problem 2: Baby names (Operations on Data Frames)

Here we load the baby names data for you. This dataset comprising counts of the given names of babies born in California each year from 1990 - 2018.

```
myfile <- "https://raw.githubusercontent.com/xjw1001001/xjw1001001.github.io/main/data/homework%20data/baby_names.csv"
data = read.csv(myfile)
head(data)
```

```
  State Sex Year      Name Count
1    CA   F 1990   Jessica  6635
2    CA   F 1990    Ashley  4537
3    CA   F 1990 Stephanie  4001
4    CA   F 1990    Amanda  3856
5    CA   F 1990  Jennifer  3611
6    CA   F 1990 Elizabeth  3170
```

For this problem, use the dataframe output or function output is enough. No need to answer in sentences. If the problem ask you of several operations, use pipes and only output the last result. If the resulting dataframe has more than 6 rows, use `head()` to show the first 6 rows.

1. What's the dimension of `baby_names`? How many observations (rows) are Male and how many observations (rows) are Female?

2. Count the number of occurrences of each distinct year and then sort the result by year.

3. Select Count, Sex, Name, and Year columns, then use filter to retrieve rows of girl names in 2010 that have larger than 3000 counts.

4. Use summarize to compute the maximum count and average count of names given for each year in 1995 to 2000, You may need to filter first, and then pipe into `knitr::kable` function to generate table output. Use 3 digits in the table.

5. **(Level 2)** Display the (first) most common boy and girl names and their counts in each of the years 2005-2010.

Hint: `slice_max()` Can help you select the row with highest count. So you need pipe `slice_max(Count)` after `group_by`.

Answer it in Level 2 parts, not here

Problem 3: Diamonds visualization

The dataset diamonds in `ggplot2` package contains the prices and other attributes of almost 54,000 diamonds.

```
head(diamonds)
```

Instructions

Level 1 Parts

Level 2 parts:

P2.5: Dataset name `data` in Problem 2.

Problem 4: Baby name visualization

Include the person you coop with:

Appendix

```
# A tibble: 6 × 10
  carat cut       color clarity depth table price     x     y     z
  <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

1. Use histogram to explore the distribution of `carat`. Choose a reasonable large binwidth (around 0.5) and another reasonable small binwidth (around 0.05). What do you find comparing these two plots? Do you discover anything unusual or surprising?

- Instruction:
    - `ggplot(?, aes(x = ?)) +`
    - `geom_?(?) +`
    - `labs(x = ?, y = ?, title = ?)`

2. What's the sample correlation between the `price` and `carat`? Hint: You can access the vector using `$`.

3. Produce a scatter plot and a hex plot between price and carat. You can try `geom_hex(bins=25)` or another value. Why hex plot is better here?

- Instruction:
    - `ggplot(?, aes(x = ?, y = ?)) +`
    - `geom_?(?) +`
    - `labs(x = ?, y = ?, title = ?)`

4. Create a horizontal boxplot to visualize the distribution of price by cut in the diamonds dataset. Choose appropriate x, y labels and title. Use `fill = "lightblue", color = "blue"` in `geom_boxplot()`. What do you find in this plot?

- Instruction:
    - `ggplot(?, aes(x = ?, y = ?)) +`
    - `geom_boxplot(fill = "lightblue", color = "blue) +`
    - `labs(x = ?, y = ?, title = ?)`

5. Create a density plot of carat facet by color in the diamond dataset. Comment any unusual information you find in the plot.

- Instruction:
    - `ggplot(?, aes(x = ?, fill = color)) +`
    - `geom_density(? = ?) +`
    - `facet_?(?, ? = ?) +`
    - `labs(x = ?, y = ?, title = ?)`

Level 2 parts:

P2.5: Dataset name `data` in Problem 2.

Display the (first) most common boy and girl names and their counts in each of the years 2005-2010.

Hint: `slice_max()` Can help you select the row with highest count. So you need pipe `slice_max(Count)` after `group_by`.

Instructions

Level 1 Parts

Level 2 parts:

P2.5: Dataset name `data` in Problem 2.

Problem 4: Baby name visualization

Include the person you coop with:

Appendix

Problem 4: Baby name visualization

1. Bar plot: We can create a bar plot showing the total count of male vs. female names in the dataset by year from 2008 to 2015.

- Instructions
  - Start from the `data %>%` . We need to obtain the dataset of total count of male and female names for year from 2008 to 2015.
  - `group_by(What variables?) %>%`
  - `summarise(total=which_function(what_variable) %>%`
  - `filter(some_condition) %>%`
  - `ggplot(aes(x = ?, y = ?, fill = ?)) +`
  - Remember concat the tidyverse functions using pipes `%>%` or `|>` , concat the ggplot layers we use `+`
  - `geom_bar(stat = ?, position = ?) +`
  - `labs` to change the labels, title
  - use some function to change the theme.

2. Line chart: We can create a line plot showing the trend in the number of names: Jessica Emily Isabella Amanda Jennifer for all these years.

- Instructions
  - First define: `names = c("Jessica", "Emily", "Isabella", "Amanda", "Jennifer")`
  - Start from the `data %>%` We need to obtain the dataset of count of the selected names for each year.
  - `filter(What_condition?) %>%`
  - `filter(Sex == "F") %>%` Why do we need to add this line? What happened if we don't add this line, can you explain why?
  - `ggplot(aes(? = ?, ? = ?, ? = ?)) +`
  - `geom_?() +`
  - `labs` to change the labels, title
  - `scale_color_viridis_d(option = "H")`
  - use some function to change the theme.

3. Word cloud: We can create a word cloud showing the most common names overall. This can give us a visual representation of the most popular names in the dataset.

- Instructions
  - Start from the `data %>%` . We need to obtain the total count of the name, sex pair.
  - `group_by(What_variables?) %>%`
  - `summarise(Count = sum(?)) %>%`
  - `arrange(desc(Count)) %>% head(50) %>%` We need this because we want only show the result of the 50 most popular names.
  - `ggplot(aes(label = ?, size = ?, color = ?)) +` We want to show the names, size represent the frequency, color represent the sex.
  - `geom_text_wordcloud() +`
  - `theme_minimal()`

Reference: ggwordcloud (https://cran.r-project.org/web/packages/ggwordcloud/vignettes/ggwordcloud.html)

Include the person you coop with:

Names:

Appendix

```
sessionInfo()
```

Instructions

Level 1 Parts

Level 2 parts:

P2.5: Dataset name `data` in Problem 2.

Problem 4: Baby name visualization

Include the person you coop with:

Appendix

```
R version 4.2.3 (2023-03-15 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22621)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] ggwordcloud_0.5.0 ggthemes_4.2.4    lubridate_1.9.2   forcats_1.0.0
 [5] stringr_1.5.0     dplyr_1.1.0       purrr_1.0.1       readr_2.1.4
 [9] tidyr_1.3.0       tibble_3.2.0      ggplot2_3.4.1     tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.10      bslib_0.4.2      compiler_4.2.3   pillar_1.8.1
 [5] jquerylib_0.1.4  tools_4.2.3      digest_0.6.31    timechange_0.2.0
 [9] jsonlite_1.8.4   evaluate_0.20    lifecycle_1.0.3  gtable_0.3.2
[13] png_0.1-8        pkgconfig_2.0.3  rlang_1.1.0      cli_3.6.0
[17] rstudioapi_0.14  yaml_2.3.7       xfun_0.37        fastmap_1.1.1
[21] withr_2.5.0      knitr_1.42       hms_1.1.3        generics_0.1.3
[25] sass_0.4.5       vctrs_0.6.0      grid_4.2.3       tidyselect_1.2.0
[29] glue_1.6.2       R6_2.5.1         fansi_1.0.4      rmarkdown_2.20
[33] tzdb_0.3.0       magrittr_2.0.3   scales_1.2.1     htmltools_0.5.4
[37] colorspace_2.1-0 utf8_1.2.3       stringi_1.7.12   munsell_0.5.0
[41] cachem_1.0.7
```