Fall 2020 STAT 240 Practice Midterm Solutions

Due N/A

Preliminaries

This practice exam aims to provide you with an example of the format of the midterm exam. Our midterm exam is scheduled for Friday, October 16, 2020 (12:00 AM CT - 11:59 PM CT). You will have 24 hours to complete the exam, and your solutions should be uploaded to Canvas by 11:59 PM CT (the time zone in Madison, WI). Note that the content of our actual midterm exam will cover material from the beginning of the semester through week 6.

A few additional things to keep in mind about the actual midterm exam:

- You are not allowed to communicate with anyone using any means (email, phone, text, social media, online discussion platforms, etc.) except the instructors of this course. You are allowed to use materials from the course and the internet. Before taking the midterm, you will need to agree to following an honor code policy.

- If you have questions during the exam, plan to post your questions on a *private* post on Piazza. To do this, select the "Individual Student(s) / Instructor(s)" option next to "Post to:" when creating your post.

- While you have until 11:59 PM CT to submit your exam, it is recommended that you begin the exam as soon as possible and read over it to see if you have any questions. You can expect for questions to be addressed during normal working hours in Madison, WI (9 AM CT - 5 PM CT). Questions posted outside that window *may* still be addressed if possible

Data

The following data files are need to complete this exam: `Police_Incident_Reports.csv` and `nfl-passing-2019-weeks-1-6.csv`.

Problems

Problem 1 (2 points)

Which of the following is **not** an aesthetic which may be set to a variable inside of `aes()` that affects the appearance of a point plotted using `geom_point()`: (a) `alpha` (b) `color` (c) `jitter` (d) `x` (e) none of the above

(c) 'jitter' does not affect the appearance of a point plotted using geom_point()

Problem 2 (2 points)

Give an example of an invalid name for an R object, and explain why it is invalid.

3object, this is an invalid name for an R object since R object cannot be started with number.

Problem 3 (2 points)

Data sets x and y each have a column named zip. Data set x has 100 rows while data set y has 1000 rows. The mutating join function xy <- mystery_join(x,y) results in a data frame xy with 78 rows. Which command is mystery_join()?

anti_join()

Problem 4 (2 points)

Briefly explain when to use geom_bar() and when to use geom_col() when making bar graphs in ggplot2.

While geom_bar() makes the height of the bar proportional to the number of cases in each group, geom_col() makes stats come in pairs.

Police data

The questions in this section use the data file Police_Incident_Reports.csv.

Problem 5 (5 points)

Read in the data set Police_Incident_Reports.csv and call this data frame police. Adjust the data frame so that IncidentDate appears in the first column (and the other variables are included in the same order).

Print the first three incidents from 2015.

REPLACE THIS TEXT WITH YOUR RESPONSE

```
police <- read_csv("C:/stat_240/data/Police_Incident_Reports.csv") %>%
  select(IncidentDate, everything())

police %>%
  filter(year(IncidentDate) == '2015') %>%
  slice_min(order_by = IncidentDate, n=3)
```

```
## # A tibble: 3 x 11
##   IncidentDate        IncidentID IncidentType      CaseNumber  Suspect  Arrested
##   <dttm>                   <dbl> <chr>             <chr>       <chr>    <chr>
## 1 2015-01-01 00:55:00      16922 Disturbance       2015-000075  <NA>    <NA>
## 2 2015-01-01 12:35:00      16924 Suspicious Person 2015-542    "Male, ~ <NA>
## 3 2015-01-02 19:56:00      16925 Robbery           2015-001928 "Black ~ <NA>
## # ... with 5 more variables: Address <chr>, Victim <chr>, Details <chr>,
## #   ReleasedBy <chr>, DateModified <dttm>
```

Problem 6 (5 points)

What were the three most common types of incidents that occurred in May, June, July, or August?

```
common_types <- read_csv("C:/stat_240/data/Police_Incident_Reports.csv") %>%
  filter(month(IncidentDate) %in% c(5:8)) %>%
  select(IncidentType) %>%
  count(IncidentType) %>%
  slice_max(order_by = n, n=3)
common_types
```

```
## # A tibble: 3 x 2
##   IncidentType          n
##   <chr>             <int>
## 1 Robbery             786
## 2 Weapons Violation   651
## 3 Battery             338
```

Problem 7 (5 points)

Which `IncidentType` values do **not** appear on Tuesdays, Wednesdays, or Thursdays in your `police` data frame?

REPLACE THIS TEXT WITH YOUR RESPONSE

```
types <- police %>%
  select(IncidentType) %>%
  distinct()
#The entire incidenttypes


types_by_wday <- police %>%
  filter(wday(IncidentDate) %in% c(3:5)) %>%
  select(IncidentType) %>%
  distinct()


anti_join(types, types_by_wday, by = "IncidentType")
```

```
## # A tibble: 10 x 1
##    IncidentType
##    <chr>
##  1 Preserve the Peace
##  2 Lost Property
##  3 Private Property Parking Complaint
##  4 Emergency
##  5 Liquor Law Violation
##  6 Towed Vehicle
##  7 Stolen Bicycle
##  8 Annoying Phone Call
##  9 Fight In Progress
## 10 Civil Dispute
```

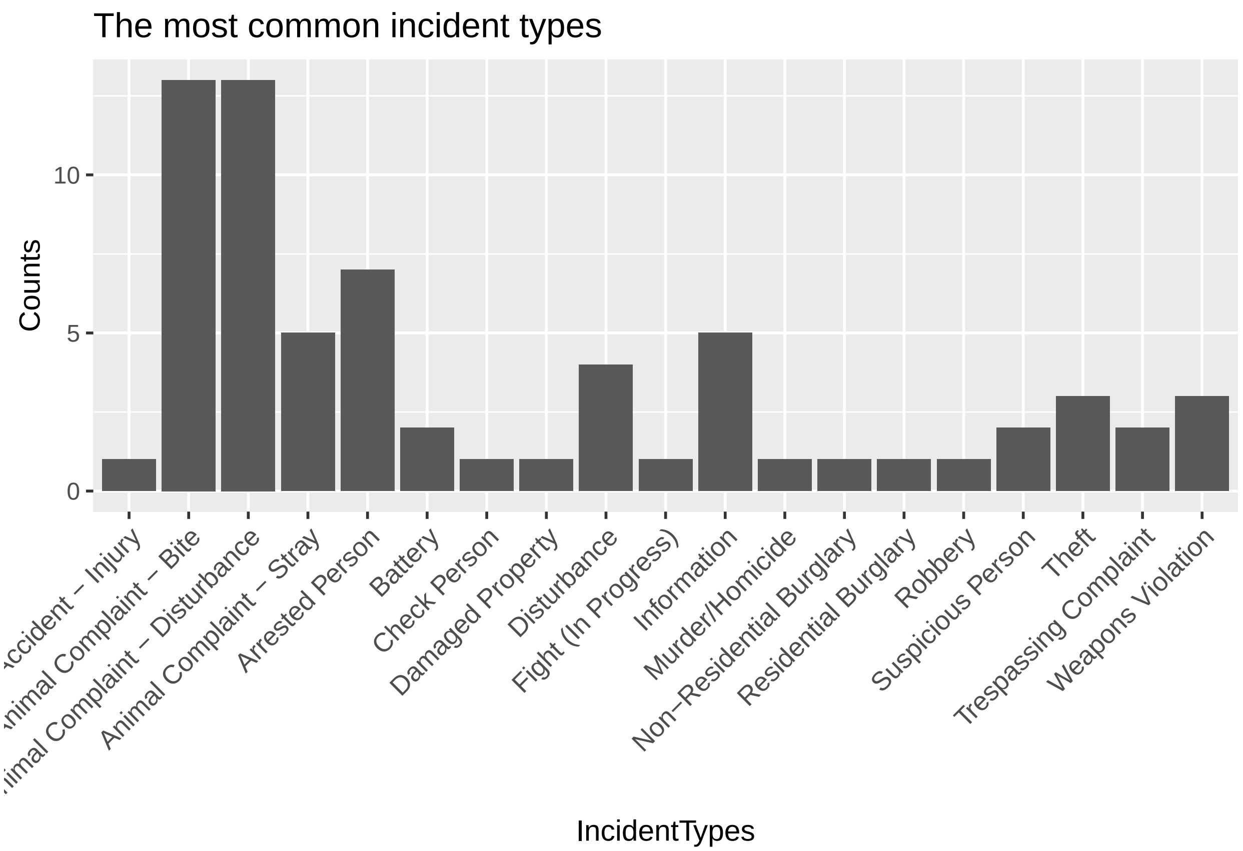
Problem 8 (5 points)

The `police` data set includes a variable `Details`. Find all incidents that include the word ‘animal’ ignoring the case of the letters. Using the resulting incidents, produce a bar plot of the incident types. Adjust the x-axis text so the incident types do not overlap. Add appropriate x and y axis labels and a descriptive title.

What is the most common incident type of the resulting incidents? (You can use the graphic or code to determine this.)

REPLACE THIS TEXT WITH YOUR RESPONSE

```
police %>%
  filter(str_detect(Details, "(?i)animal")) %>%
  select(IncidentType) %>%
  count(IncidentType) %>%
  distinct() %>%
  ggplot(aes(x = IncidentType, y = n)) +
  geom_bar(stat = "identity") +
  xlab("IncidentTypes") +
  ylab("Counts") +
  ggtitle("The most common incident types") +
  theme(axis.text.x = element_text(hjust = 1, size = 10, angle = 45))
```



NFL data

The questions in this section use the data file `nfl-passing-2019-weeks-1-6.csv`.

Problem 9 (5 points)

Read in the data set `nfl-passing-2019-weeks-1-6.csv` and call this data frame `nfl`. The data contains information about players (primarily quarterbacks) and their performances during weeks 1 through 6 of the 2019 NFL season.

`Result` indicates the outcome of the game and includes the score of the game in the format of "Vising team score" - "Home team score". `Location` indicates if the player was playing in their home field (`home`) or not (`away`). `Cmp%` is the completion percentage of passes attempted by the player throughout the game; a pass is completed if a player on the passer's team catches the ball.

Add a variable called `outcome` that takes the value `won` if the player's team won the game, `lost` if the player's team lost the game, and `tie` if the game ended in a tie. Add another variable called `points` that has the score for each player's team.

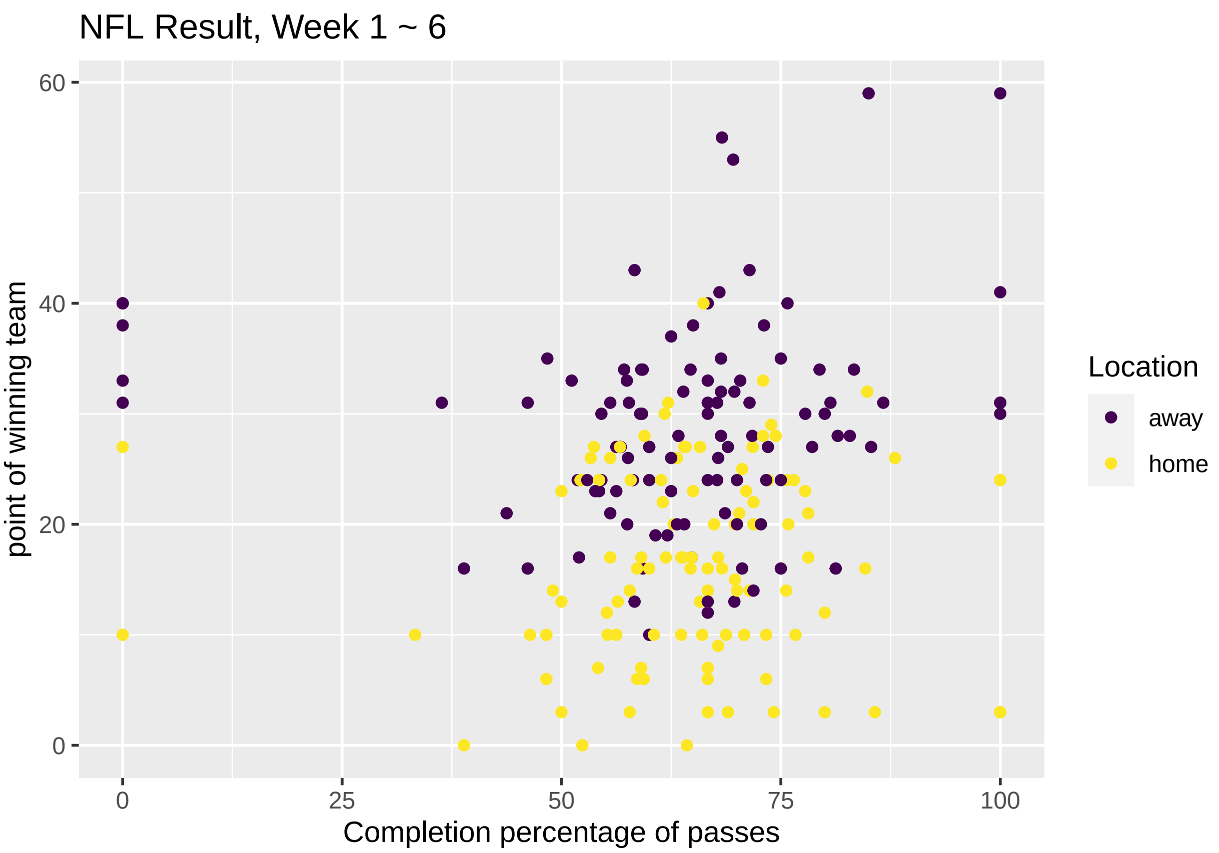
Create a scatter plot of `points` (y-axis) versus completion percentage (x-axis). Color the points by `Location`. Add appropriate x and y axis labels and a descriptive title.

REPLACE THIS TEXT WITH YOUR RESPONSE

```r
nfl <- read_csv("C:/stat_240/data/nfl-passing-2019-weeks-1-6.csv")

nfl2 <- nfl %>%
  mutate(outcome = case_when(str_detect(Result, "W") ~ "won",
                            str_detect(Result, "L") ~ "lost",
                            str_detect(Result, "T") ~ "tie")) %>%
  rename(cmp_percent = "Cmp%") %>%
  mutate(home = str_sub(Result, 2, -1)) %>%
  separate(home, c("away", "home"), sep = "-") %>%
  mutate(points = case_when(outcome == "won" & Location == "home" ~ home,
                           outcome == "won" & Location == "away" ~ away,
                           outcome == "lost" & Location == "home" ~ away,
                           outcome == "lost" & Location == "away" ~ home,
                           outcome == "tie" ~ away))


ggplot(nfl2, aes(x = cmp_percent, y = as.numeric(points), color = Location)) +
  geom_point() +
  xlab("Completion percentage of passes") +
  ylab("point of winning team") +
  ggtitle("NFL Result, Week 1 ~ 6")
```



Problem 10 (5 points)

Att indicates the number of pass attempts in the game, and Yds indicates the yards passing for the player

Among players whose team won the game, the location of the game was away, and the player attempted 20 or more passes in the game, about what percentage of these players had fewer than 200 yards passing?

REPLACE THIS TEXT WITH YOUR RESPONSE

```
nfl2 %>%
  filter(outcome == "won" & Location == "away" & Att >= 20) %>%
  mutate(yds = Yds < 200) %>%
  select(yds, Yds) %>%
  summarise(percent = mean(yds) * 100)
```

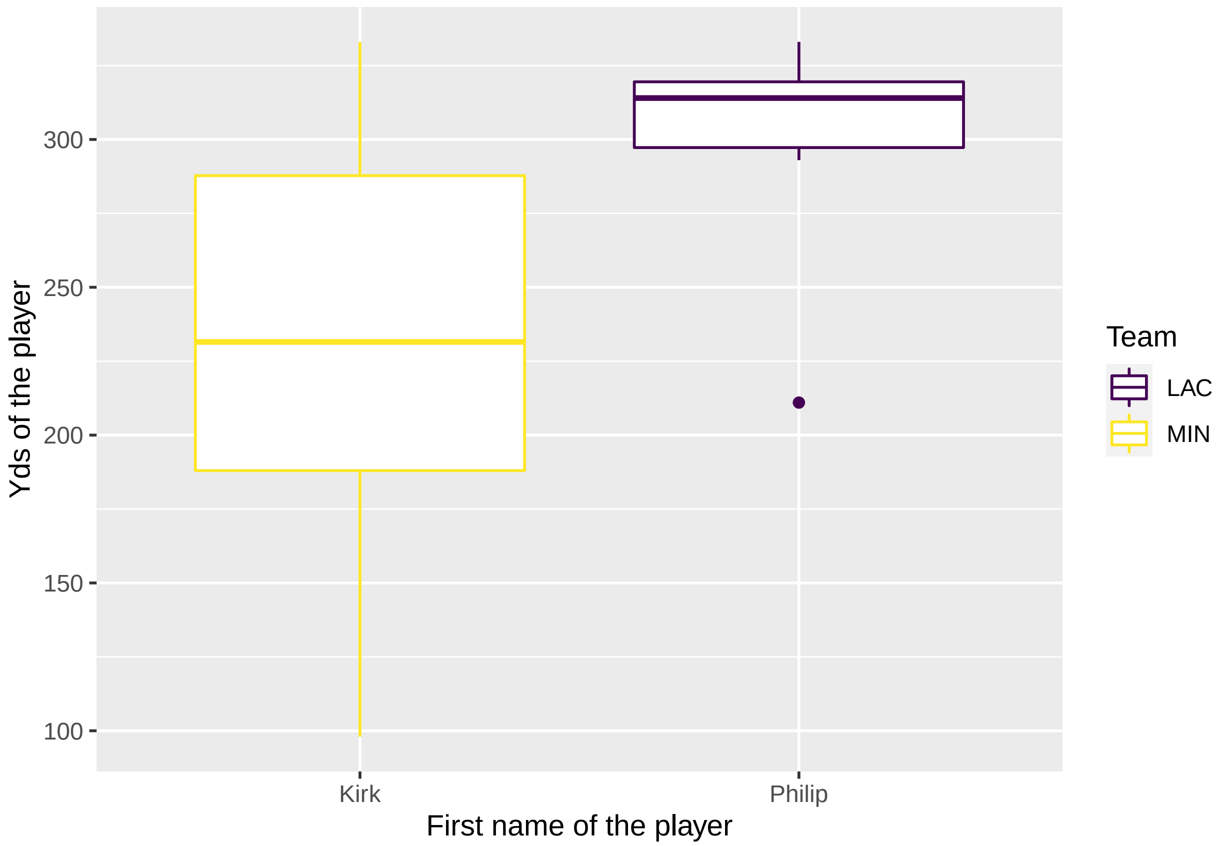
```
## # A tibble: 1 x 1
##   percent
##     <dbl>
## 1      26
```

Problem 11 (5 points)

Player has the name and unique ID for each player, and Tm indicates the player's team. Create side-by-side boxplots of Yds for the players with a first name that begins and ends with the same letter. Color the points by `Tm`. Add appropriate x and y axis labels and a descriptive title. Change the legend title to "Team."

REPLACE THIS TEXT WITH YOUR RESPONSE

```r
nfl2 %>%
  separate(Player, into = c("Name", "ID"), sep = "\\\\") %>%
  separate(Name, into = c("First", "Last"), sep = " ") %>%
  filter(str_detect(First, "(?i)^([a-z]).*\\1$")) %>%
  ggplot(aes(x = First, y = Yds, color = Tm)) +
  geom_boxplot() +
  xlab("First name of the player") +
  ylab("Yds of the player") +
  guides(color = guide_legend(title = "Team"))
```



Submission

Once you have completed all of the questions, knit the R Markdown document to create an HTML file. To submit this Exam, go to our Canvas site and select "Assignments" on the left panel, and upload both the edited .Rmd and HTML files to the place designated for the exam.