Assignment 7

Due Friday, October 23, 11:59 PM CT Problems 1 - 6 use the exoplanet data set to practice using string and regular expression commands. The data were pulled from the NASA Exoplanet Archive on September 3, 2020 (with 4,276 exoplanets confirmed as of this date, but this data set includes planet *candidates* as well).

The following block of code will read in the exoplanet data from the file *exoplanets-3sept2020.csv*, filter to include only confirmed exoplanets, select and rename a subset of variables, and add an index with the row number.

```
planets <- read_csv("C:/stat_240/data/exoplanets-3sept2020.csv") %>%
  filter(default_flag == 1) %>%
  select(pl_name, hostname, discoverymethod, disc_year, disc_facility,
         sy_pnum, pl_rade, pl_bmasse, rowupdate) %>%
  rename(planet=pl_name, star=hostname, method=discoverymethod,
         facility=disc_facility, year=disc_year, number=sy_pnum,
         radius=pl_rade, mass=pl_bmasse, update=rowupdate) %>%
  mutate(index = row_number()) %>%
  select(index, everything())
```

Problems

1

Create and display a table which shows the ten most frequently occurring facilities with the name "Observatory" in the title, arranged from most to least. Which facility is fifth on the list and how many times does it appear?

```
freq_obs <- planets %>%
  mutate(condition = str_detect(facility, "Observatory")) %>%
  filter(condition == TRUE) %>%
  count(facility) %>%
  slice_max(n, n=10)
freq_obs
```

```
## # A tibble: 10 x 2
##    facility                                       n
##    <chr>                                      <int>
##  1 La Silla Observatory                         245
##  2 W. M. Keck Observatory                       175
##  3 Haute-Provence Observatory                    51
##  4 Lick Observatory                              32
##  5 Las Campanas Observatory                      29
##  6 McDonald Observatory                          29
##  7 Paranal Observatory                           25
##  8 Okayama Astrophysical Observatory             23
##  9 Roque de los Muchachos Observatory            22
## 10 Bohyunsan Optical Astronomical Observatory    17
```

Las Campanas Obervatory is fifth on the list and it appears 29 times.

2

One of the stars has the name "2MASS J04414489+2301513". Create a regular expression which matches only this string. Display the regular expression and then the string expression in R used to represent this regular expression. (Put single back ticks around your answers so that they appear properly after knitting.) Finally, write an R expression using the command `str_replace(string,` pattern, replacement) which takes “2MASS J04414489+2301513” as the input string and uses appropriate values for pattern and replacement so that the string that is returned is the string representation of the regular expression you found earlier.

```
^2.*3$
```

```
## Change to eval = TRUE in chunk arguments before knitting


str_replace(string = "2MASS J04414489+2301513",
            pattern = "^2.*3$",
            replacement = "2MASS J04414489+2301513")
```

```
## [1] "2MASS J04414489+2301513"
```

3

The convention to name most planets appears to be the name of the star followed by a space and a suffix, usually a single letter such as ‘b’. Find all exoplanets where the name of the planet does not begin with the name of the star followed by a space. Note: special care is required if the name of the star contains a symbol with special meaning in a regular expression, such as +. Create a data frame with the rows where the planet does not follow this convention and select the columns index, planet, and star. Modify the strings in planet and star by changing spaces ’ ’ to underscores ’_’ and adding a slash ‘/’ at the start and end of each string. Display this modified data frame. How many such planets do not follow the naming convention? (*Hint: You may find it helpful to create a column with the regular expression you wish to compare to the planet name.*)

```
planets3 <- planets %>%
  mutate(star = str_replace_all(star,"\\+"," "),planet = str_replace_all(planet,"\\+"," "))%>%
  filter(!str_detect(planet,str_c(star," ")))%>%
  select(index,planet,star)%>%
  mutate(star = str_replace_all(star,"\\s","_"),planet = str_replace_all(planet,"\\s","_"))%>%
  mutate(star = str_c("/",star,'/'),planet = str_c("/",planet,'/'))
planets3
```

```
## # A tibble: 27 x 3
##    index planet                         star
##    <int> <chr>                          <chr>
##  1    11 /2MASS_J01033563-5515561_AB_b/ /2MASS_J01033563-5515561_A/
##  2   116 /EPIC_201615463_c/             /K2-166/
##  3   117 /EPIC_201754305_d/             /K2-16/
##  4   118 /EPIC_201833600_c/             /K2-50/
##  5   120 /EPIC_205950854_c/             /K2-168/
##  6   134 /EPIC_220554210_c/             /K2-282/
##  7   237 /GJ_9066_b/                    /GJ_83.1/
##  8   238 /GJ_9066_c/                    /GJ_83.1/
##  9   696 /HD_21749_c/                   /GJ_143/
## 10  1408 /KIC_8540376_b/                /KOI-7892/
## # ... with 17 more rows
```

27 planets do not follow the naming convention.

4

Find all the planets where the planet name is the name of the star, a space, and then something else we will label a suffix. Create a data frame where you add a variable named suffix which contains this suffix. In how many cases is the suffix a single lower case letter? Create and display a summary table that counts the number of times each single lower case letter is used, arranged from most to least frequent use.

```r
planets4 <- planets %>%
  mutate(star = str_replace_all(star,"\\+"," "),planet = str_replace_all(planet,"\\+"," "))%>%
  filter(str_detect(planet,str_c(star," ")))%>%
  mutate(suffix = word(planet, -1, sep = star))%>%
  select(index,planet,star,suffix)%>%
  filter(!str_length(suffix)>2)%>%
  filter(str_detect(suffix,"[:lower:]"))%>%
  group_by(suffix)%>%
  count(suffix)%>%
  distinct()

planets4
```

```
## # A tibble: 7 x 2
## # Groups:   suffix [7]
##   suffix     n
##   <chr>  <int>
## 1 " b"    3161
## 2 " c"     710
## 3 " d"     239
## 4 " e"      89
## 5 " f"      31
## 6 " g"      11
## 7 " h"       5
```

4246 Cases.

5

Create a data frame that contains all exoplanets where the planet name begins with the name of the star followed by a space and a suffix which is not a single lower case letter. Reduce this data frame to the columns index, planet, star, and suffix and display it. How many such planets are there? *(Note: Such planets might be orbiting binary star systems where the planet naming convention is different to indicate which star the planet orbits.)*

```r
planets5 <- planets %>%
  mutate(star = str_replace_all(star,"\\+"," "),planet = str_replace_all(planet,"\\+"," "))%>%
  filter(str_detect(planet,str_c(star," ")))%>%
  mutate(suffix = stringr::word(planet, -1, sep = star))%>%
  select(index,planet,star,suffix)%>%
  filter(str_detect(suffix,"[:upper:]")|str_length(suffix)>2)
planets5
```

```
## # A tibble: 3 x 4
##   index planet      star   suffix
##   <int> <chr>       <chr>  <chr>
## 1   166 GJ 229 A c  GJ 229 " A c"
## 2  1003 HU Aqr AB b HU Aqr " AB b"
## 3  1004 HU Aqr AB c HU Aqr " AB c"
```

There are three planets.

6

The column `update` in the exoplanet data set has character values where some entries have a date only, such as `5/14/14` and others have a date and time, such as `9/4/18 16:14`. Create a new column named `update_format` with the value "date" if the format is like `5/14/14`, "datetime" if the format is like `9/4/18 16:14`, and "other" if it is something else. Count how many rows have each type.

```
exoplanet6 <- planets %>%
  mutate(update_format = case_when(str_detect(update,"^\\d{1,2}/\\d{1,2}/\\d{1,2}$") ~"date",str_detect
  group_by(update_format)%>%
  count(update_format)%>%
  distinct()

exoplanet6
```

```
## # A tibble: 2 x 2
## # Groups:   update_format [2]
##   update_format     n
##   <chr>         <int>
## 1 date           1376
## 2 datetime       2900
```

REPLACE THIS TEXT WITH YOUR RESPONSE

Probability Problems

A discrete random variable X has possible values and probabilities contained in the following data frame which includes some missing values. Use this distribution for the following four problems.

```
prob7 = tibble(
  x = c(1,2,3,5),
  p = c(0.1, 0.4, NA, 0.2)
)
```

7

What is $P(X = 3)$? Briefly explain how you arrived at the answer.

Since the sum of the probability should be 1, P(X=3) = NA = 1-(0.1+0.4+0.2) = 0.3 Therefore, P(X=3) = 0.3

8

What are the expected value (mean) and variance of the distribution of X? (Do the calculations in R.)

```
Mean <- 1*0.1 + 2*0.4 + 3*0.3 + 5*0.2
Mean
```

```
## [1] 2.8
```

```
Variance = 1^2*0.1 + 2^2*0.4 + 3^2*0.3 + 5^2*0.2  - {Mean}^2
Variance
```

```
## [1] 1.56
```

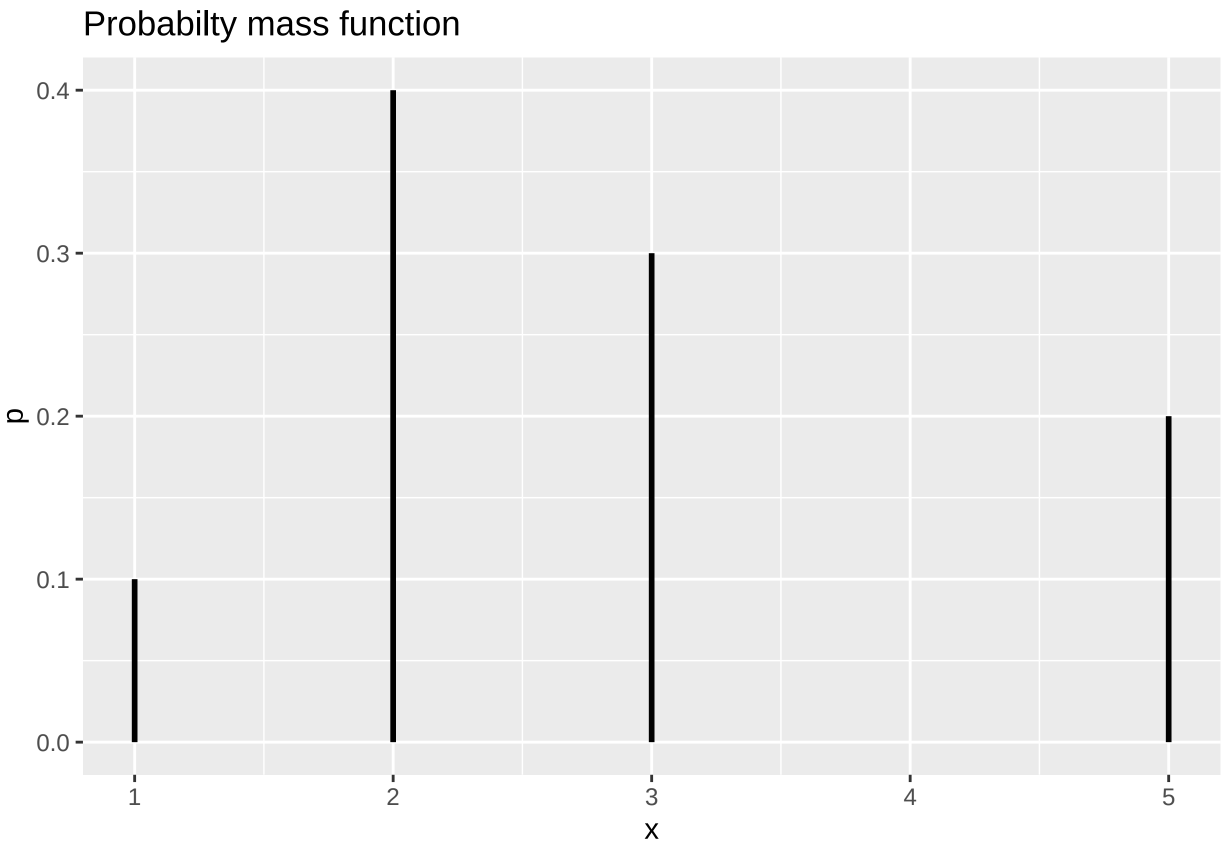
Mean: 2.8 Variance: 1.56

9

Create a graph which has a line segment from 0 to the probability for each possible value of X to show its probability mass function. *(Hint: see how to use the* **ggplot2** *command* `geom_segment()`*.)*

```
prob7 = tibble(
  x = c(1,2,3,5),
  p = c(0.1, 0.4, 0.3, 0.2)
)

ggplot(prob7, mapping = aes(x=x, y=p))+
  geom_segment(aes(xend = x, yend=rep(0,1)), size=1)+
  ggtitle("Probabilty mass function")
```

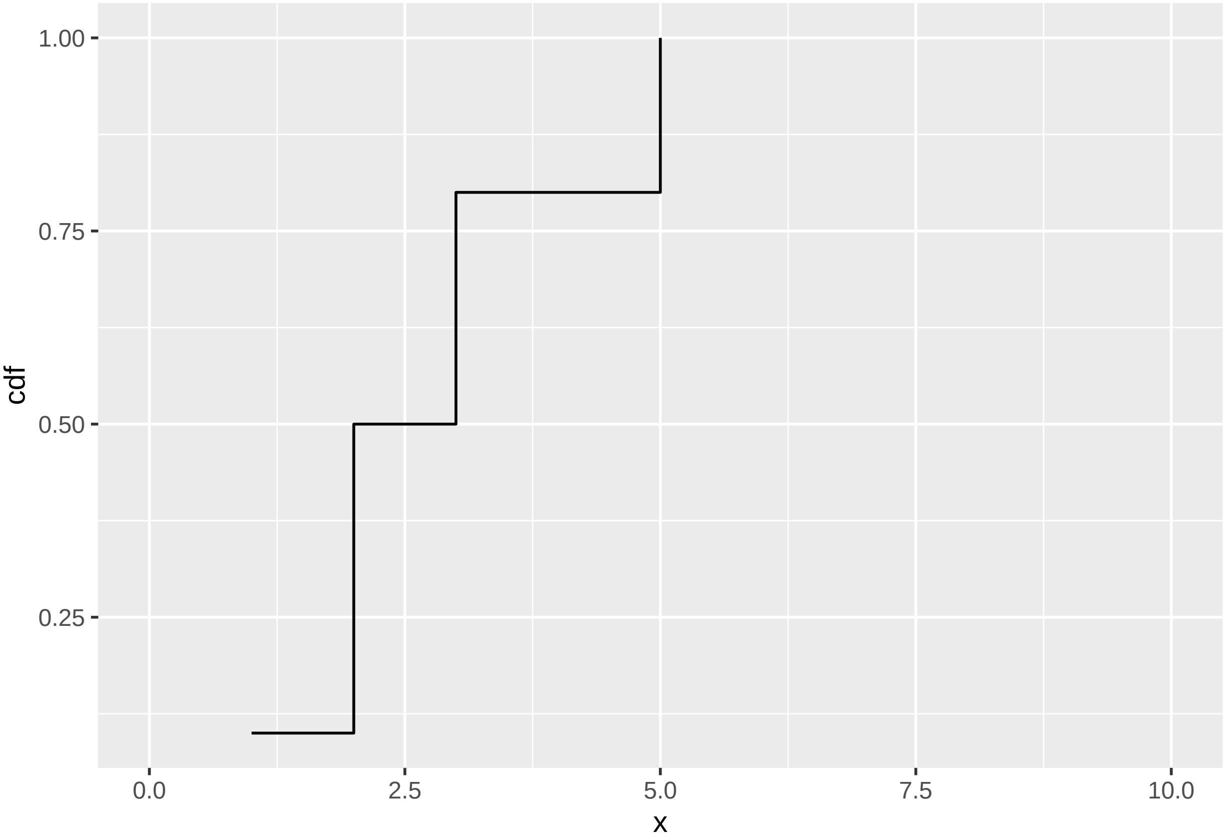


10

Add a column named `cdf` which contains the value $P(X \leq x)$ for each case. Use the function `geom_step()` to graph this function. The graph of the function will look better if you also include x values below 1 and above 5. *(Hint: the base R function* `cumsum()` *which calculates a cumulative sum may be helpful.)*

```
prob7 <- tibble(
  x = c(1,2,3,5),
  p = c(0.1, 0.4, 0.3, 0.2),
  cdf = cumsum(p)
)
```

```
ggplot(prob7, mapping = aes(x=x, y=cdf)) +
  geom_step()+
  xlim(c(0,10))
```


cdf
x