Assignment 4

Kyle Yeo

Due Friday, September 25, 11:59 PM

Problems

1

> Calculate the minimum, 25th percentile, mean, median, and maximum value of mass for planets that were discovered using the method `Radial Velocity` or the `Transit` method in the data set. Display these summary statistics separately for each method.

```
Radial <- planets %>%
  filter(method == 'Radial Velocity') %>%
  group_by(method) %>%
  select(mass) %>%
  drop_na() %>%
  summarise(percentile = quantile(mass, .25), mean = mean(mass), median = median(mass), max = max(mass))
Radial
```

```
## # A tibble: 1 x 5
##   method          percentile  mean median    max
##   <chr>                <dbl> <dbl>  <dbl>  <dbl>
## 1 Radial Velocity       23.9  891.   372. 17668.
```

```
Transit <- planets %>%
  filter(method == 'Transit') %>%
  group_by(method) %>%
  select(mass) %>%
  drop_na() %>%
  summarise(percentile = quantile(mass, .25), mean = mean(mass), median = median(mass), max = max(mass))
Transit
```

```
## # A tibble: 1 x 5
##   method  percentile  mean median   max
##   <chr>        <dbl> <dbl>  <dbl> <dbl>
## 1 Transit       15.4  406.   178. 8654.
```

Do most of these planets have an estimated mass less than, greater than, or about the same as the mass of the Earth?

Most of these planets have an estimated mass largely greater than the same as the mass of the Earth.

2

> Count the number of exoplanets that have been discovered that have a mass less than or equal to the mass of the Earth, and display the count and minimum and maximum mass of these planets. Similarly, count the number of exoplanets that have been discovered that have a radius less than or equal to the radius of the Earth, and display the count and minimum and maximum radius of these planets.

```r
exo_mass <- planets %>%
  group_by(planet) %>%
  filter(!is.na(mass)) %>%
  filter(mass <= 1) %>%
  select(mass) %>%
  summarise(n = n(), minimum =min(mass), maximum = max(mass))
exo_mass
```

```
## # A tibble: 15 x 4
##    planet             n minimum maximum
##    <chr>          <int>   <dbl>   <dbl>
##  1 GJ 9827 c          1   0.84    0.84
##  2 K2-266 c           1   0.290   0.290
##  3 Kepler-128 b       1   0.77    0.77
##  4 Kepler-128 c       1   0.9     0.9
##  5 Kepler-138 b       1   0.066   0.066
##  6 Kepler-138 d       1   0.64    0.64
##  7 KOI-55 b           1   0.44    0.44
##  8 KOI-55 c           1   0.655   0.655
##  9 PSR B1257+12 b     1   0.02    0.02
## 10 TRAPPIST-1 b       1   0.85    0.85
## 11 TRAPPIST-1 d       1   0.41    0.41
## 12 TRAPPIST-1 e       1   0.62    0.62
## 13 TRAPPIST-1 f       1   0.68    0.68
## 14 YZ Cet b           1   0.75    0.75
## 15 YZ Cet c           1   0.98    0.98
```

```r
exo_radius <- planets %>%
  group_by(planet) %>%
  filter(!is.na(radius)) %>%
  filter(radius <= 1) %>%
  select(radius) %>%
  summarise(n = n(), minimun = min(radius), maximum = max(radius))
exo_radius
```

```
## # A tibble: 163 x 4
##    planet                 n minimun maximum
##    <chr>              <int>   <dbl>   <dbl>
##  1 EPIC 201497682 b       1   0.692   0.692
##  2 EPIC 201833600 c       1   1       1
##  3 EPIC 206215704 b       1   0.9     0.9
##  4 EPIC 206317286 b       1   0.96    0.96
##  5 HD 21749 c             1   0.892   0.892
##  6 K2-116 b               1   0.69    0.69
##  7 K2-136 b               1   0.99    0.99
##  8 K2-137 b               1   0.89    0.89
##  9 K2-209 b               1   0.869   0.869
## 10 K2-210 b               1   0.819   0.819
## # ... with 153 more rows
```

3

Only a handful of planets have both an estimated mass AND an estimated radius less than those of the Earth. What are the names of these planets and what methoused to d(s) were detect them? Print a data frame that has the star name, planet name, method, mass, and radius of these planets.

```
new_planets <- planets %>%


  filter(mass < 1) %>%
  filter(radius < 1) %>%
  drop_na() %>%
  select(star, planet, method, mass, radius)
new_planets
```

```
## # A tibble: 6 x 5
##   star       planet       method                          mass radius
##   <chr>      <chr>        <chr>                          <dbl>  <dbl>
## 1 K2-266     K2-266 c     Transit                        0.290  0.705
## 2 KOI-55     KOI-55 b     Orbital Brightness Modulation 0.44   0.759
## 3 KOI-55     KOI-55 c     Orbital Brightness Modulation 0.655  0.867
## 4 Kepler-138 Kepler-138 b Transit                        0.066  0.522
## 5 TRAPPIST-1 TRAPPIST-1 d Transit                        0.41   0.772
## 6 TRAPPIST-1 TRAPPIST-1 e Transit                        0.62   0.918
```

4

What are the planet names and estimated masses of **all** the detected planets orbiting the host stars from the previous questions? That is, for all host stars that have at least one planet with an estimate mass AND an estimated radius less than or equal to those of the Earth, what are the names and masses of all their orbiting planets. You may find it useful to use the command `pull(star)` to extract the column of star names from the previous question. Arrange these planets from most massive to the least massive.

```
new_stars <- new_planets %>%
  pull(star)

orbiting_planets <- planets %>%
  filter(star %in% new_stars) %>%
  filter(!is.na(mass)) %>%
  select(planet, mass) %>%
  arrange(desc(mass))
orbiting_planets
```

```
## # A tibble: 15 x 2
##    planet         mass
##    <chr>         <dbl>
##  1 K2-266 e     14.3
##  2 K2-266 b     11.3
##  3 K2-266 d      8.9
##  4 Kepler-138 c  1.97
##  5 TRAPPIST-1 c  1.38
##  6 TRAPPIST-1 g  1.34
##  7 TRAPPIST-1 b  0.85
##  8 TRAPPIST-1 f  0.68
##  9 KOI-55 c      0.655
## 10 Kepler-138 d  0.64
## 11 TRAPPIST-1 e  0.62
## 12 KOI-55 b      0.44
## 13 TRAPPIST-1 d  0.41
```

```
## 14 K2-266 c      0.290
## 15 Kepler-138 b  0.066
```

```r
#?order_by
```

5

> Which stars hosts the three most massive planet? Display the star name, planet name, method, year, and mass, and add a new variable called `mass_j` that contains the mass in units of Jupiter Mass.
> Note: 1 Jupiter Mass = 317.8 Earth Mass (approximately)

```r
planets %>%
  select(star, planet, method, year, mass) %>%
  distinct() %>%
  slice_max(mass,n=3) %>%
  mutate(mass_j = mass/317.8)
```

```
## # A tibble: 3 x 6
##   star       planet       method           year   mass mass_j
##   <chr>      <chr>        <chr>           <dbl>  <dbl>  <dbl>
## 1 BD+20 2457 BD+20 2457 b Radial Velocity  2009 17668.   55.6
## 2 HD 148284  HD 148284 b  Radial Velocity  2018 10711.   33.7
## 3 HR 2562    HR 2562 b    Imaging          2016  9535.   30.0
```

What is the mass (in Jupiter Mass) of the most massive exoplanet, what year and by which method was it detected?

The mass of the most massive exoplanet is 55.59525 and it was detected by 'Radial Velocity' in 2009.

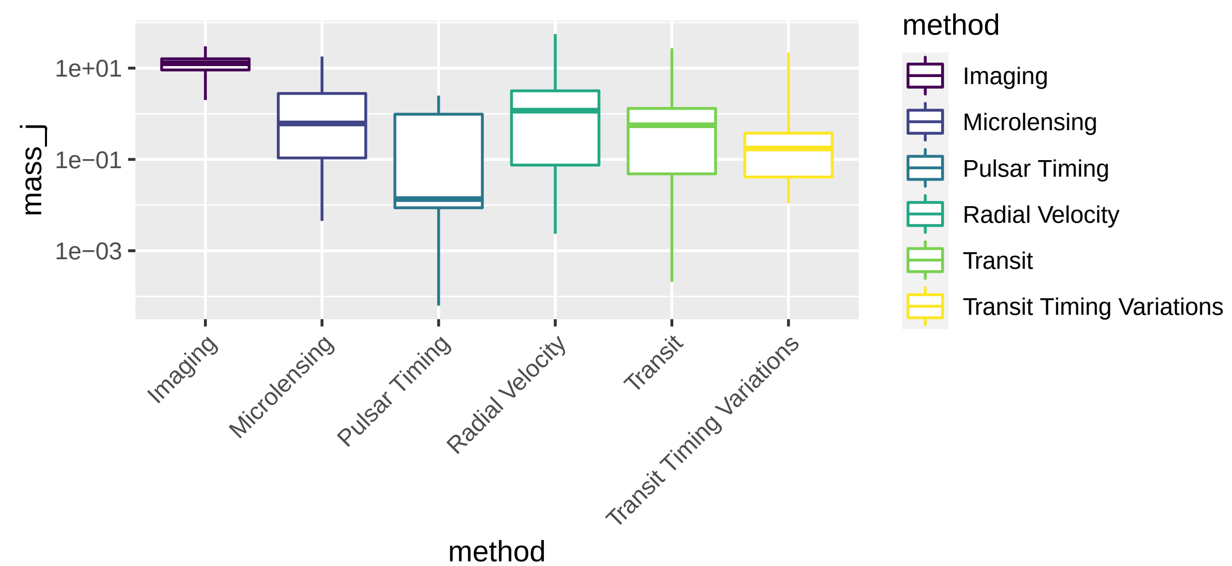
6

> Create a graph with side-by-side boxplots to compare the distribution of estimated planet mass in Jupiter Mass units by detection method. Remove the planets that were detected using `Astrometry`, `Disk Kinematics`, `Eclipse Timing Variations`, `Orbital Brightness Modulation`, `Pulsar Timing`, or `Pulsar Timing Variations`. Also, remove all values with missing masses; be careful here not remove observations with *any* missing value…you only want to exclude those with missing masses. In the `geom_boxplot`, set `coef=Inf` (create the plot with and without this setting to see what it does). Color the boxplots by method and put the y-axis on the `log10` scale using the `trans` option in `scale_y_continuous()`. The horizontal axis labels may overlap a bit; add this line of code to your ggplot: `theme(axis.text.x = element_text(angle = 45, hjust=1))` and decide if it helps. Try swiching the value for `angle` and see what happens.

```r
q6_planets <- planets %>%
  filter(method != "Astrometry") %>%
  filter(method != "Disk Kinematics") %>%
  filter(method != "Eclipse Timing Variations") %>%
  filter(method != "Orbital Brightness Modulation") %>%
  filter(method != "Pulsation Timing") %>%
  filter(method != "Pulsation Timing Variations") %>%
  filter(!is.na(mass)) %>%
  mutate(mass_j = mass/317.8)
#q6_planets

ggplot(q6_planets, aes(x=method, y=mass_j, color = method)) +
  geom_boxplot(coef=Inf) +
```

```
  scale_y_continuous(trans = "log10") +
  theme(axis.text.x = element_text(angle = 45, hjust =1))
```



What does the `coef=Inf` do in the box plot function?

remove all outliers

What does changing the value for `angle` do?

It differentiates the angle that x-variables have with

From this graphic, does it seem there are differences in the ability of methods to detect exoplanets with different masses?

yes, there are differences in the ability of methods to detect exoplanets with different masses.

7

> Make a data frame that excludes exoplanets discovered by the transit method and includes the variables `planet`, `method`, `year`, `radius`, and `mass`. Remove any of the remaining observations with missing values. Use this data frame to create a scatterplot with discovery year on the x-axis and radius on the y-axis. Plot the symbol color according to discovery method and have the size of the points be according to the estimated mass. Use the command `scale_size_binned()` to adjust the symbol size `breaks` to be the minimum, 25th percentile, median, 75th percentile, and maximum of the new data frame's mass estimates. See `?scale_size_binned()` for more details. Use `labs(color = "Method", size = "Mass")` to adjust the legend labels.

```
df <- planets %>%
  filter(method != "Transit") %>%
  select(planet, method, year, radius, mass) %>%
  drop_na()


breaks <- df %>%
  summarise(min = min(mass), quan = quantile(mass,.25), median = median(mass), quan2 = quantile(mass, .
```

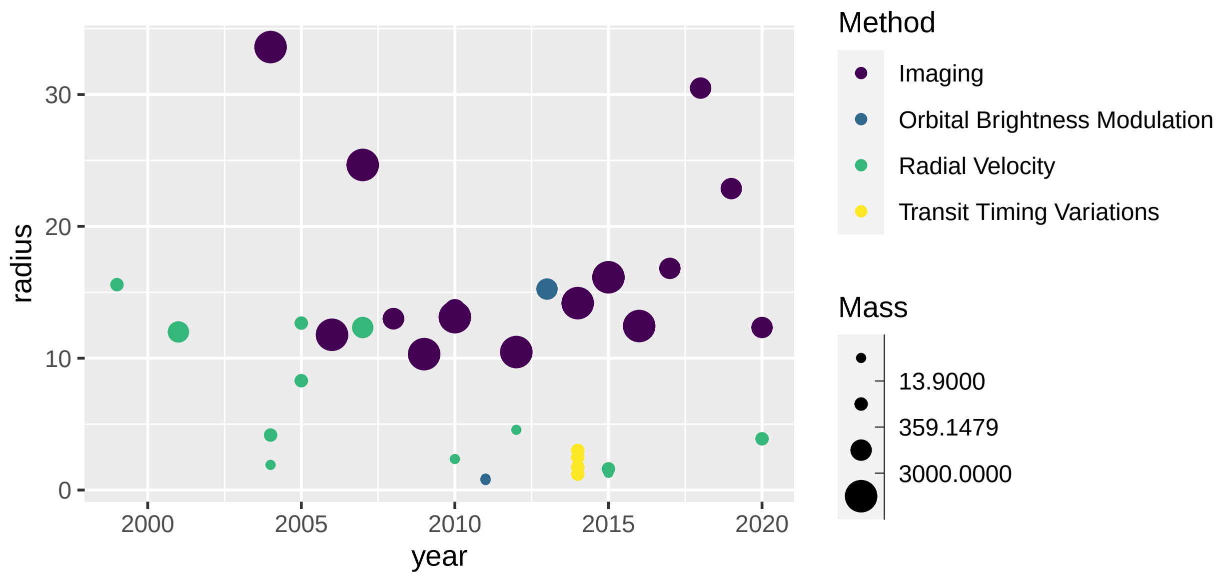
```r
visual <- ggplot(df, mapping = aes(x = year, y = radius)) +
  geom_point(aes(color = method, size = mass)) +
  labs(size = "Mass", color = "Method")







visual + scale_size_binned(breaks = breaks)
```



What patterns do you notice in this graphic? Explain.

planets that were detected by imaging method tend to be more massive planets. Also, the planets that were detected recently are less massive.

8

Create a variable called `decade` that assigns the observations to the appropriate decade between 1980 and 2020. Set the labels to be characters showing the range of years for the decade such as "1981-1990", "1991-2000", etc. You may find the `cut()` command useful here. Then make side-by-side box plots of mass by decade. Add appropriate titles to the x-axis and y-axis and put the y-axis on the `log10` scale using the `trans` option in `scale_y_continuous()`.

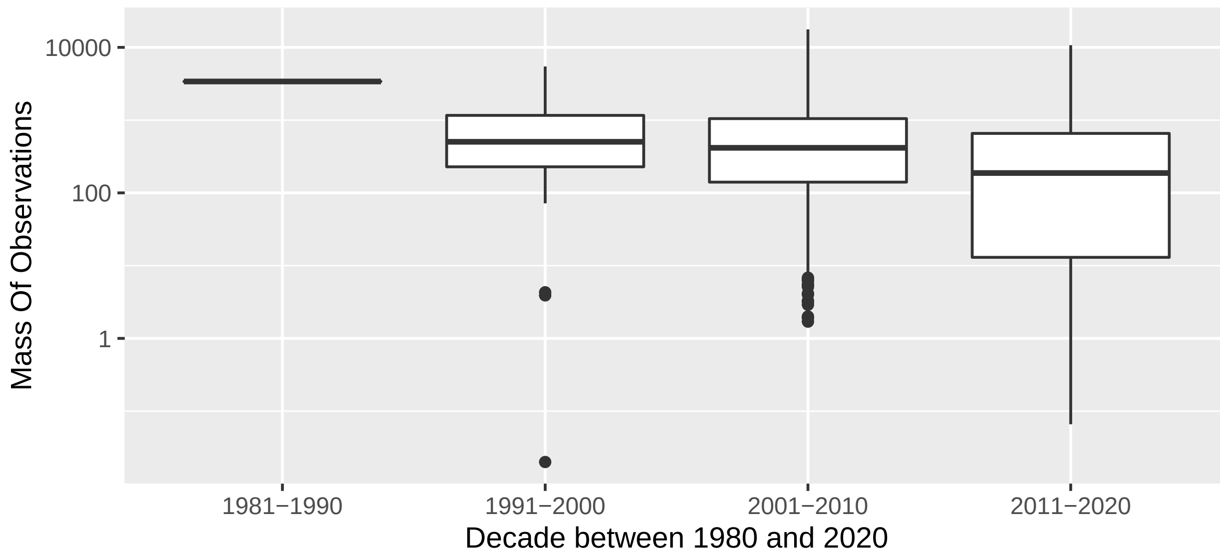
```r
q8_planets <- planets
x <- q8_planets$year
decade <- cut(x, breaks = c(1980,1990,2000,2010,2020), labels = c("1981-1990", "1991-2000", "2001-2010"
q8_planets %>%
  mutate(decade = decade) %>%
  select(mass, decade)
```

```
## # A tibble: 4,276 x 2
##      mass decade
##     <dbl> <fct>
##  1 6166. 2001-2010
##  2 4685. 2001-2010
##  3 1526. 2001-2010
```

```
## 4 1481. 2001-2010
## 5  566. 1991-2000
## 6 3274. 2001-2010
## 7 3000  2001-2010
## 8  289. 2011-2020
## 9  632. 2001-2010
## 10  273. 2001-2010
## # ... with 4,266 more rows
```

```r
ggplot(q8_planets) +
  geom_boxplot(aes(x=decade, y=mass)) +
  scale_y_continuous(trans = "log10") +
  xlab("Decade between 1980 and 2020") +
  ylab("Mass Of Observations")
```

```
## Warning: Removed 2513 rows containing non-finite values (stat_boxplot).
```



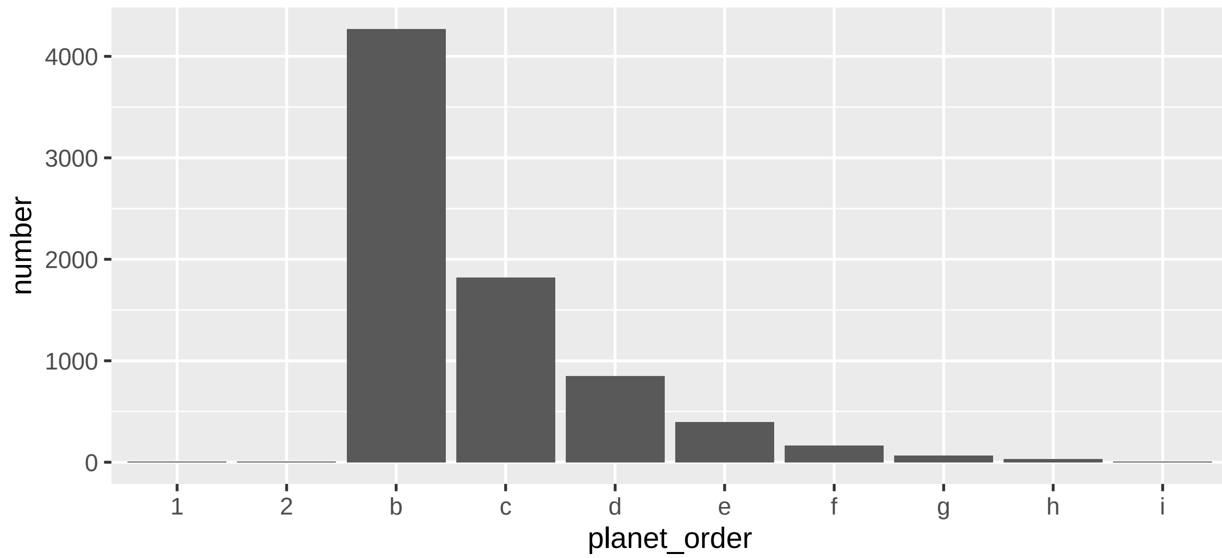
What sort of trend is present across decades in terms of the median mass of discovered exoplanets?

The median mass of discovered exoplanets is steadily decreasing over decades.

9

The naming convention used for planets is that the first planet discovered has a name that ends with "b" (often the name of the host star followed by the "b"). If a second planet is discovered it will use "c", then "d", etc. Let's use this convention to find the distribution of planets. Create a new variable called `planet_order` that pulls the last value of `planet`. To get the last value, we can use the R package stringr's command `str_sub()`: `str_sub(planet,-1)` (the first input specifies the variable and the -1 grabs the first value from the end). Then create a bar plot of these values.

```r
naming_convention <- planets %>%
  mutate(planet_order = str_sub(planet, -1))
ggplot(naming_convention, aes(x=planet_order, y = number))  +
  geom_bar(stat = "identity")
```




```r
naming_convention %>%
  filter(planet_order == "1" | planet_order =="2") %>%
  select(year, method, planet_order)
```

```
## # A tibble: 7 x 3
##    year method  planet_order
##   <dbl> <chr>   <chr>
## 1  2019 Transit 1
## 2  2019 Transit 2
## 3  2020 Transit 1
## 4  2020 Transit 2
## 5  2019 Transit 1
## 6  2019 Transit 1
## 7  2019 Transit 2
```

There are some planets that do not appear to follow the naming convention. What years were these planets discovered and by which method? What naming convention is used for these planets?

In 2019, there were five planets that do not appear to follow the naming convention, all of which were detected by transit method and in 2020, there were just two planets that do not appear to follow the naming convention, which also were detected by transit method.