Homework 1

Name: Nuttapong Mekvipad, Ryan William Moreau, Rani Nielsen, Silvija Pupsaite, Liuqing Zheng

Group: 7

Question 1

Before we perform any data analysis further, we need to look at the data set first to understand the structure of data and how to plan our analysis.

```
library(babynames)
```

```
## Warning: package 'babynames' was built under R version 3.4.4
```

```
head(babynames)
```

```
## # A tibble: 6 x 5
##    year sex   name          n   prop
##   <dbl> <chr> <chr>     <int>  <dbl>
## 1  1880 F     Mary       7065 0.0724
## 2  1880 F     Anna       2604 0.0267
## 3  1880 F     Emma       2003 0.0205
## 4  1880 F     Elizabeth  1939 0.0199
## 5  1880 F     Minnie     1746 0.0179
## 6  1880 F     Margaret   1578 0.0162
```

a) List the top 5 female baby names starting with P, regardless of year, as a table.

The code for question 1a) is seen in the chunk below. We first filter for the rows with the female names starting with 'P'. Then we summarise the total number of names and use head function to obtain the top 5 female baby names starting with P.

The most popular female baby names starting with P was Patricia followed by Pamela, Phyllis, Peggy and Paula respectively.

```
filtered_baby <- babynames %>% filter(str_detect(name, "^P"), sex=='F') %>%
  group_by(name) %>% summarise(number.of.baby = sum(n)) %>% arrange(desc(number.of.baby))

head(filtered_baby, 5)
```

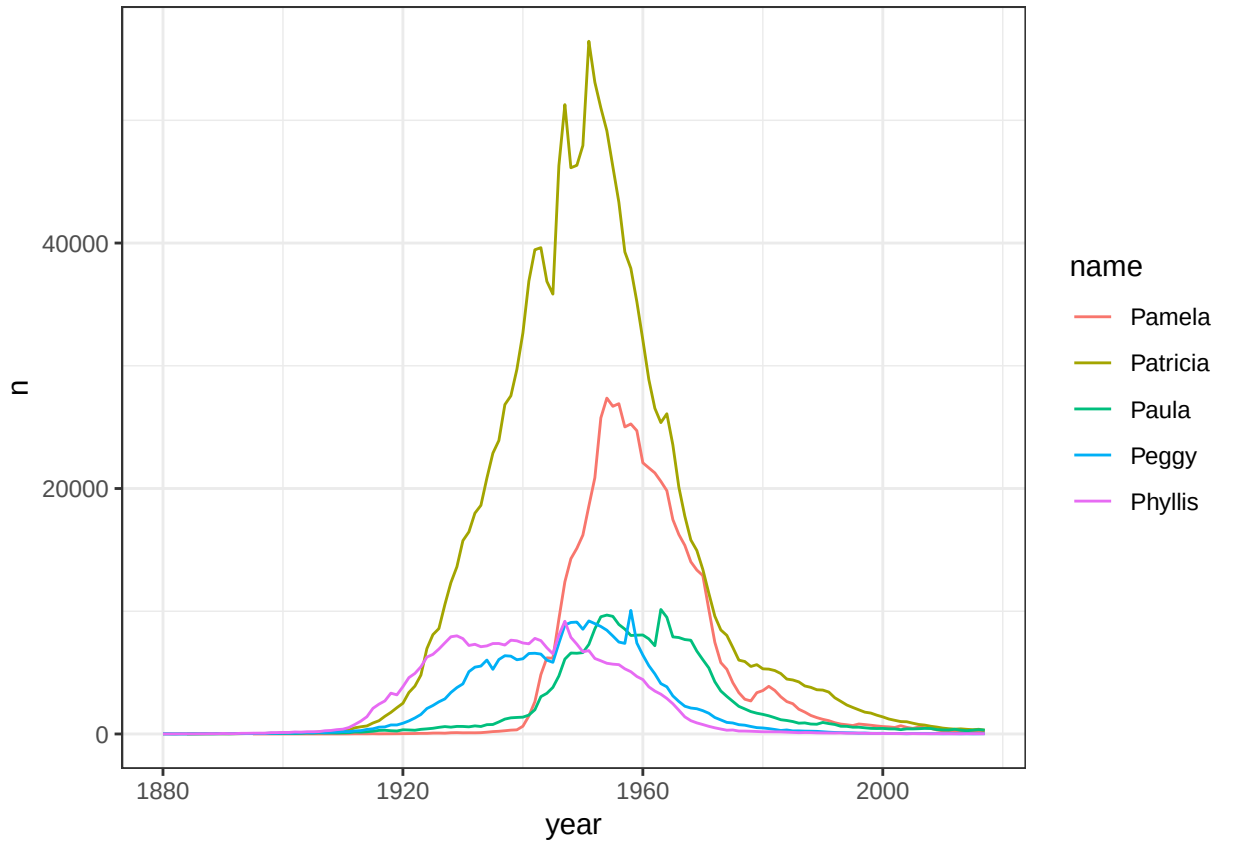
```
## # A tibble: 5 x 2
##   name     number.of.baby
##   <chr>             <int>
## 1 Patricia        1571692
## 2 Pamela           594174
## 3 Phyllis          322369
## 4 Peggy            292585
## 5 Paula            278003
```

b) Using the results from a, plot their occurrences as a function of year using a line plot. Comment on your results. If you get strange results, explain them and/or improve the plot

We plot of the occurences of each name over the year to see the change in popularity of that name. From line plot we can see that all of the name in top 5 reach the peak around mid of 20th century and then start to decline around that time.

```
top5_name <- filtered_baby$name[1:5]

babynames %>% filter(name %in% top5_name, sex=='F') %>% ggplot(aes(x=year, y=n, color=name)) +
  geom_line() + theme_bw()
```



Question 2

To see whether there is any difference in how common the babies named Arwen were between year 2004 and 1990, we need to compare the proportion of Arwen's born in 1990 with proportion of Arwen's born in 2004. To compare the proportion between those two years we chose the chi-squre test, beacuse two of our factors were a catagorical variable (the year and the status of being Arwen or not).

To perform the test, we first create contigency table of year vs name.

```
n_total <- babynames %>% filter(year==2004 | year == 1990) %>% group_by(year) %>%
  summarise(n.total=sum(n))
n_arwen <- babynames %>% filter(name=="Arwen", year==2004 | year == 1990) %>%
  transmute(year = year, n.arwen = n)

contig.table.arwen <- inner_join(n_arwen, n_total) %>%
  transmute(year = year, n.arwen = n.arwen, n.not.arwen = n.total - n.arwen, prop.arwen =
              n.arwen/n.total)
```

```
## Joining, by = "year"
```

```
contig.table.arwen
```

```
## # A tibble: 2 x 4
```

```
##    year n.arwen n.not.arwen prop.arwen
##   <dbl>   <int>       <int>      <dbl>
## 1  1990      10     3950982 0.00000253
## 2  2004     166     3818195 0.0000435
```

We can see that the proportion of Arwen in 2004 was about 20-folds higher than in 1990. However, we need to test whether this difference in proportion is statistically significant. So we applied a chi-square test over the contigency table.

The null hypothesis of chi-square test is that the proportion of Arwen in year 1990 is not different from proportion of Arwen in year 2004. The result of chi-square test is below.

```
chisq.test(as.matrix(contig.table.arwen[,2:3]))
```

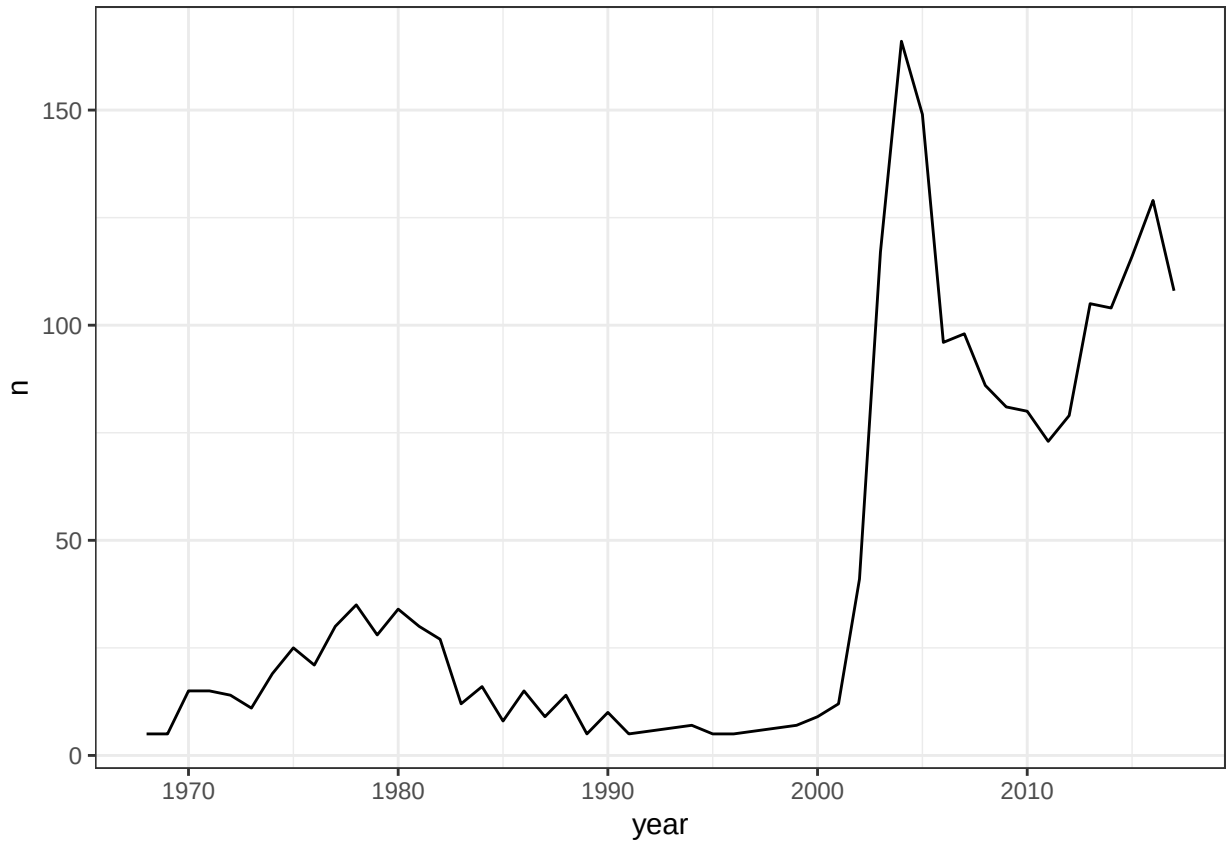
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  as.matrix(contig.table.arwen[, 2:3])
## X-squared = 141.89, df = 1, p-value < 2.2e-16
```

From the result we can see that the p-value is lower than 2.2e-16 which is lower that our threshold at 0.05. Therefore, we rejected null hypothesis. This means that the proportion of Arwen in 2004 is significantly different from in 1990.

Therefore, we may say that the propotion of Arwen in 2004 was higher than in 1990 as showed in contigency table above, and this difference (proportion of Arwen being higher in 2004) was significant when tested with chi-square test.

Interestingly, we can see from graph below that the occurences of Arwen greatly increased around 2001 and reached the peak at 2004. This coincided with premiere of Lord of the Rings trilogy which have main character named Arwen. This might be one of the reasons why the proportion of Arwen in 2004 was different from 1990.

```
babynames %>% filter(name=="Arwen") %>% ggplot(aes(x=year, y=n)) + geom_line() + theme_bw()
```



Question 3

```
flowers <- read_tsv('flowers.txt')
```

```
## Parsed with column specification:
## cols(
##   Sepal.Length = col_double(),
##   Sepal.Width = col_double(),
##   Petal.Length = col_double(),
##   Petal.Width = col_double(),
##   Species = col_character()
## )
```

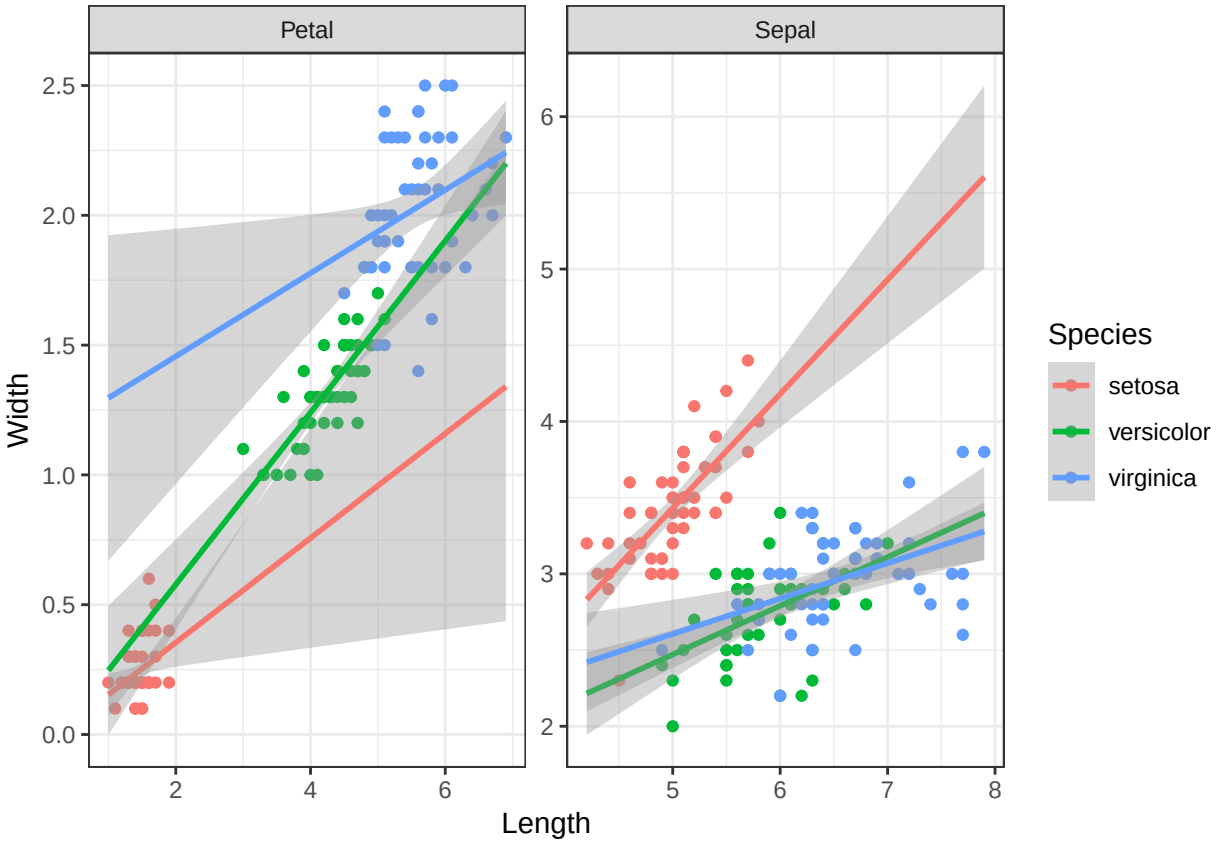
To replicate the plot, we first separated original data in to 2 tibble which were a tibble containing measurement of petal, and a tibble containing measurement of sepal. Both tibble also had the column of species name and the column that indicated the flower organ type.

We then merged 2 tibble of sepal and petal using bind_rows function, then plotted the line plot of linear model with confidence interval using geom_smooth function. We need to mention that to obtain the plot that looks exactly like in the instruction, we need to call the ggplot with geom_point and facet_wrap first to get full plot area. Then we called the geom_smooth last to overlay the line and area indicating confidence interval.

```
sepal <- flowers %>% select(Sepal.Length, Sepal.Width, Species) %>%
  transmute(Width = Sepal.Width, Length = Sepal.Length, Species = Species) %>%
  mutate(organ.type = "Sepal")
```

```
petal <- flowers %>% select(Petal.Length, Petal.Width, Species) %>%
  transmute(Width = Petal.Width, Length = Petal.Length, Species = Species) %>%
  mutate(organ.type = "Petal")

bind_rows(sepal, petal) %>% ggplot(aes(x=Length, y=Width, color = Species)) + geom_point() +
  theme_bw() + facet_wrap(~organ.type, scales = "free") +
  geom_smooth(method="lm", fullrange=TRUE)
```



Question 4

```
chip_mm5 <- read_tsv('chip_mm5.txt')
```

```
## Parsed with column specification:
## cols(
##   chr = col_character(),
##   start = col_double(),
##   end = col_double(),
##   score = col_double()
## )
```

Here we want to test two hypothesis which are: 1) Binding scores are dependent on chromosome 2) Binding site widths (end-start) are dependent on chromosome

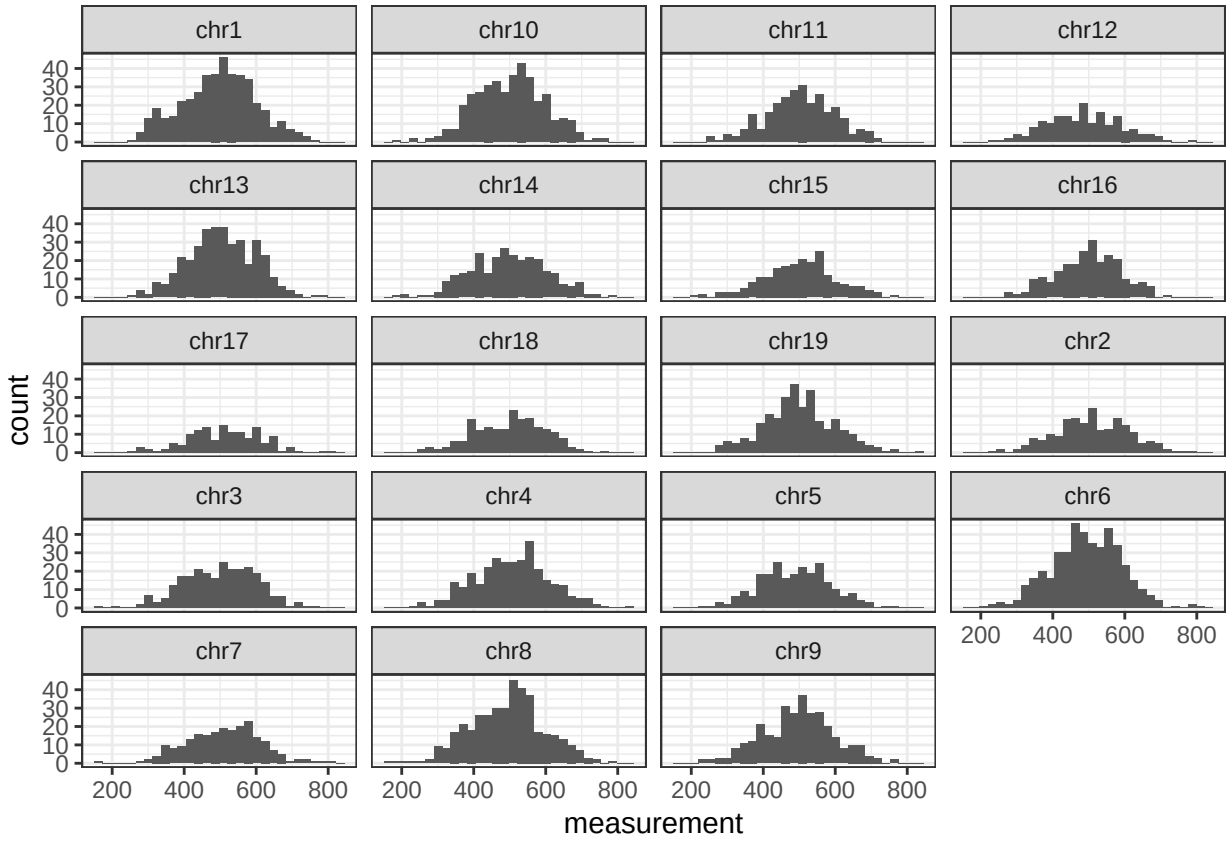
For both hypothesis, we want to test the dependency of continuous variables (score or width) on the catagorical variable (number of chromosome). To test the dependency of continuous on catagorical variable, we can use a one-way ANOVA test or Kruskal-Wallis test to see whether there is a significant difference of means score/gene width between chromosome or whether there is a location shift of core/gene width between

chromosome. If there is a significant difference in score/width or location shift, it means that there is some dependency between score/width and chromosome number.

Before we choose whether we will use one-way ANOVA test or Kruskal-Wallis test, we need to see the distribution of score and width of each chromosome.

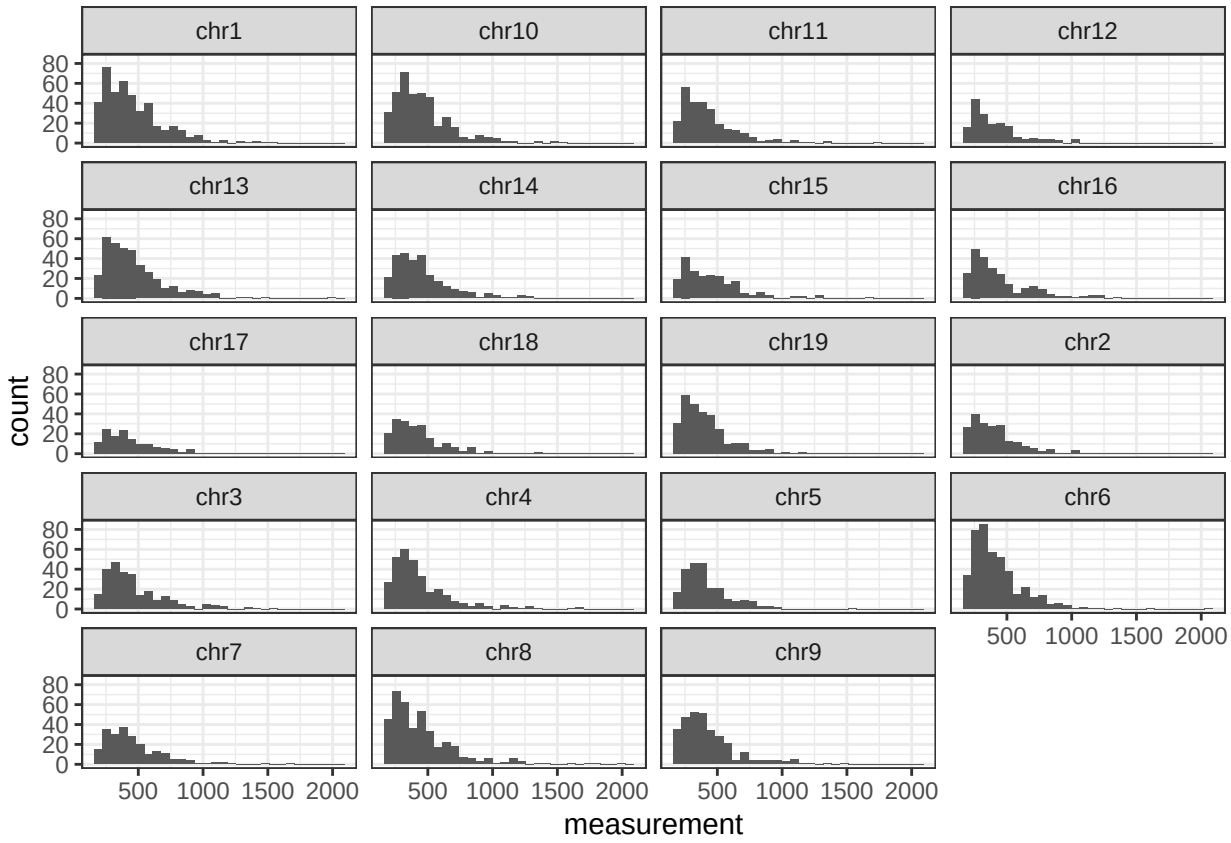
```
chip_mm5 <- chip_mm5 %>% mutate(width = abs(end-start))
chip_mm5 %>% gather(key = "type.of.measurement", value="measurement",score, width,
                    na.rm = TRUE) %>%
  filter(type.of.measurement == "score") %>%
  ggplot(aes(x=measurement)) + geom_histogram() + theme_bw() +
  facet_wrap(~chr, ncol = 4)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
chip_mm5 %>% gather(key = "type.of.measurement", value="measurement",score, width,
                    na.rm = TRUE) %>%
  filter(type.of.measurement == "width") %>%
  ggplot(aes(x=measurement)) + geom_histogram() + theme_bw() +
  facet_wrap(~chr, ncol = 4)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the plot, we can see that the shape of distribution of the score on each chromosome was similar to normal distribution, while the shape of distribution of width on each chromosome was similar to Poisson distribution. Therefore, we used a one-way ANOVA test to answer first research question, and Kruskal-Wallis test to answer the second research question.

The null hypothesis of first research question (Binding scores are dependent on chromosome or not) was the means of scores are not differnt between each chromosome.

The result below shows that p-value of one-way ANOVA test was 0.4298 which was higher than cut off at 0.05. Therefore, we retain the null hypothersis that the means of scores are not differnt between each chromosome. This means that there is no dependency between score and chromosome number.

```
chip_mm5 <- na.omit(chip_mm5)
oneway.test(score ~ chr ,data = chip_mm5)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  score and chr
## F = 1.0228, num df = 18.0, denom df = 1797.5, p-value = 0.4298
```

The null hypothesis of second research question (Binding site widths (end-start) are dependent on chromosome or not) was the location shift of gene widths has no relation to chromosome number.

The results below show that the p-value of Kruskal-Wallis test was 0.003416 which lower than 0.05. Therefore, we rejected null hypothesis that the location shift of widths has no relation to chromosome number. This means that there is a dependency of genes width on chromosome number.

```
kruskal.test(chip_mm5$width, as.factor(chip_mm5$chr))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  chip_mm5$width and as.factor(chip_mm5$chr)
## Kruskal-Wallis chi-squared = 38.411, df = 18, p-value = 0.003416
```

Next, we wanted to see which pairs of chromosome showed significant location shift of gene widths, so we peformed pairwise Wilcoxon test. From the result, we can see that only 8 pairs gave p-value lower than 1, and for the rest of pairs the p-value = 1.0. Moreover, none of the pairs which have p-value lower than 1.0 gave significant results. The reason why the pairwise test showed no significant result might be that the power to detect the statistic difference might not be enough here due to low effective size. In addition, we performed bonferroni adjustment for p-value of all pairs, so the p-values were adjusted to be higher to campensate for multiple test. Therefore, it was harder to get significant pairwise results.

However, this does not mean that there is no dependency of genes width on chromosome number as the hypothesis that we test here was that the location shift between each pair of chromosome is 0 which is different from the Kruskal-Wallis test.

```
pairwise_table <- pairwise.wilcox.test(chip_mm5$width, as.factor(chip_mm5$chr),
                                       p.adjust.methods = "bonferroni")

pvalue_table <- as.data.frame(pairwise_table$p.value)

as.tibble(rownames_to_column(pvalue_table)) %>%
  gather(key = "chr", value = "pvalue", -rowname) %>% filter(pvalue < 1)
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
## # A tibble: 8 x 3
##   rowname chr   pvalue
##   <chr>   <chr>  <dbl>
## 1 chr12   chr10  0.612
## 2 chr19   chr10  0.613
## 3 chr13   chr12  0.232
## 4 chr3    chr12  0.512
## 5 chr19   chr13  0.261
## 6 chr2    chr13  0.476
## 7 chr3    chr19  0.623
## 8 chr3    chr2   0.895
```