Business Statistics End of Term Assessment IB94X0 2023-2024

5582755

2023-12-21

This is to certify that the work I am submitting is my own. All external references and sources are clearly acknowledged and identified within the contents. I am aware of the University of Warwick regulation concerning plagiarism and collusion.

No substantial part(s) of the work submitted here has also been submitted by me in other assessments for accredited courses of study, and I acknowledge that if this has been done an appropriate reduction in the mark I might otherwise have received will be made.

Tasks

Question-1

Effect on Bike Hire as a result of COVID restrictions. We have to perform regression analysis to examine the effects of three elements of the COVID response: Work From Home, Rule of 6 Indoors, and the Eat Out to Help Out upon bike hire and effect of potential differences in the year, month, days of the week.

Question-2

Effect of reviews of the books on their sale and effect of sale price upon number of sales across different genres and answer the two Questions.

a. Effect of reviews on the sales of books.

b. Effect of sale price upon the number of sales, and difference across genres.

Required Libraries

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ─
─
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ─
─
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conf
licts to become errors
```

```
options(width=100)
#install.packages("kableExtra")
library(kableExtra) #for reported table
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
#install.packages("emmeans")
library(emmeans) #to find CI
#install.packages("gridExtra")
library(gridExtra) #for arrange graph
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
#install.packages("grid")
library(grid)
install.packages("Hmisc")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(ggplot2)
install.packages("car")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

Question 1

This report is to analyze effect of Three COVID restrictions: Work From Home, Rule of 6 Indoors, and the Eat Out to Help Out on number of bike hires in London.

Data Dictionary

This data shows number of bike hires per day in London from 2010-07-30 to 2023-09-30 and COVID restrictions during that period. The variables of the data are described in table below

Variable	Description

date	dates from 2010-07-30 to 2023-09-30
hires	No of bikes hired on the day
schools_closed	if the schools were closed on that day or not
pubs_closed	if the pubs bwere closed or not
shops_closed	if the shops were closed or open
eating_places_closed	if the eating places were closed or not
stay_at_home	if the stay at home ristriction was there
household_mixing_indoors_banned	if individuals from different households were allowed for gathering or mixing indoors
wfh	if work from home was implemented or not
rule_of_6_indoors	if the indoor gatherings were limited to 6 people
curfew	if there was curfew on the day
eat_out_to_help_out	people were encouraged to dine out
day	day of the week
month	month of the year
year	from year 2010 to 2023

Read Data

```
#first we read data
data_bikes <- read_csv("London_COVID_bikes.csv")
```

```
## Rows: 4812 Columns: 15
## ── Column specification ────────────────────────────────────────────────────
──────────────────────
## Delimiter: ","
## chr   (2): day, month
## dbl  (12): Hires, schools_closed, pubs_closed, shops_closed, eating_places_clos
ed, stay_at_home,...
## date  (1): date
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this messag
e.
```

```
#we select the three data columns of COVID restrictions: Work From Home, Rule of 6 Indoors, and the Eat Out to Help Out scheme which we are interested in.
Covidres_data <- data_bikes[, c("date", "Hires", "wfh", "rule_of_6_indoors", "eat_out_to_help_out", "day", "month", "year")]
```

Data Quality Checks

```
#check data summary
str(Covidres_data)
```

```
## tibble [4,812 × 8] (S3: tbl_df/tbl/data.frame)
##  $ date               : Date[1:4812], format: "2010-07-30" "2010-07-31" "2010-08-01" ...
##  $ Hires              : num [1:4812] 6897 5564 4303 6642 7966 ...
##  $ wfh                : num [1:4812] 0 0 0 0 0 0 0 0 0 0 ...
##  $ rule_of_6_indoors  : num [1:4812] 0 0 0 0 0 0 0 0 0 0 ...
##  $ eat_out_to_help_out: num [1:4812] 0 0 0 0 0 0 0 0 0 0 ...
##  $ day                : chr [1:4812] "Fri" "Sat" "Sun" "Mon" ...
##  $ month              : chr [1:4812] "Jul" "Jul" "Aug" "Aug" ...
##  $ year               : num [1:4812] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
```

```
#check for duplicates
sum(duplicated(Covidres_data))
```

```
## [1] 0
```

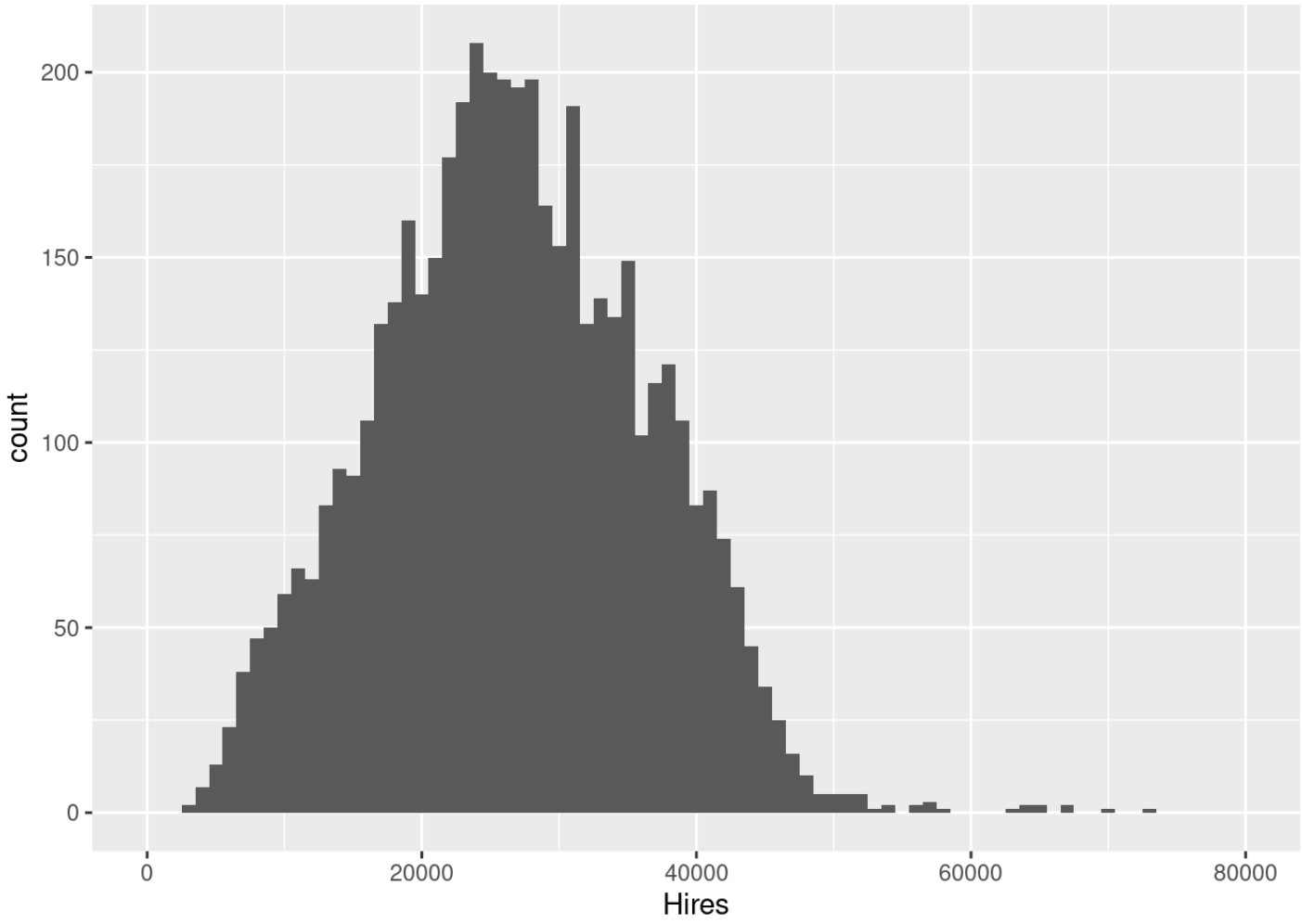
```
#check for missing values
sum(is.na(Covidres_data))
```

```
## [1] 0
```

```
# No missing values or duplicated data.
```

```
## Simple histogram showing the overall distribution and checking for outliers.
ggplot(Covidres_data) + geom_histogram(aes(Hires), binwidth = 1000) + xlim(0, 80000)
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```




```
# The distribution density is normally distributed, without any extreme high value
s that could potentially be designated as outliers.
```

```
#We covert the month, year and day into facors.
Covidres_data$year <- as.factor(Covidres_data$year)
Covidres_data$month <- factor(Covidres_data$month, levels = c("Jan","Feb","Mar","A
pr","Jun","Jul","Aug","Sep","Oct","Nov","Dec"))
Covidres_data$day <- factor(Covidres_data$day, levels = c("Mon","Tue","Wed","Th
u","Fri","Sat","Sun"))

# we Check for how many time were these restrictions imposed within the given data
set from year 2010 - 2023
count(Covidres_data, wfh)
```

```
## # A tibble: 2 × 2
##     wfh     n
##   <dbl> <int>
## 1     0  3718
## 2     1  1094
```

```
count(Covidres_data, rule_of_6_indoors)
```

```
## # A tibble: 2 × 2
##   rule_of_6_indoors     n
##               <dbl> <int>
## 1                 0  4716
## 2                 1    96
```

```
count(Covidres_data, eat_out_to_help_out)
```

```
## # A tibble: 2 × 2
##   eat_out_to_help_out     n
##                 <dbl> <int>
## 1                   0  4784
## 2                   1    28
```

#Visualise Number of Bike hires in London from year 2010 - 2023

```
# Simple histogram showing the overall distribution of Bike Hires in London
p.hires.distribution <- ggplot(Covidres_data,aes(x=Hires,..density..)) +
  geom_histogram(colour = "pink") +
  geom_density(col="red") +
  labs(caption = "Figure 1: Distribution of bike hires in london from 2010-2023")

p.hires.distribution
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was gener
ated.
```

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

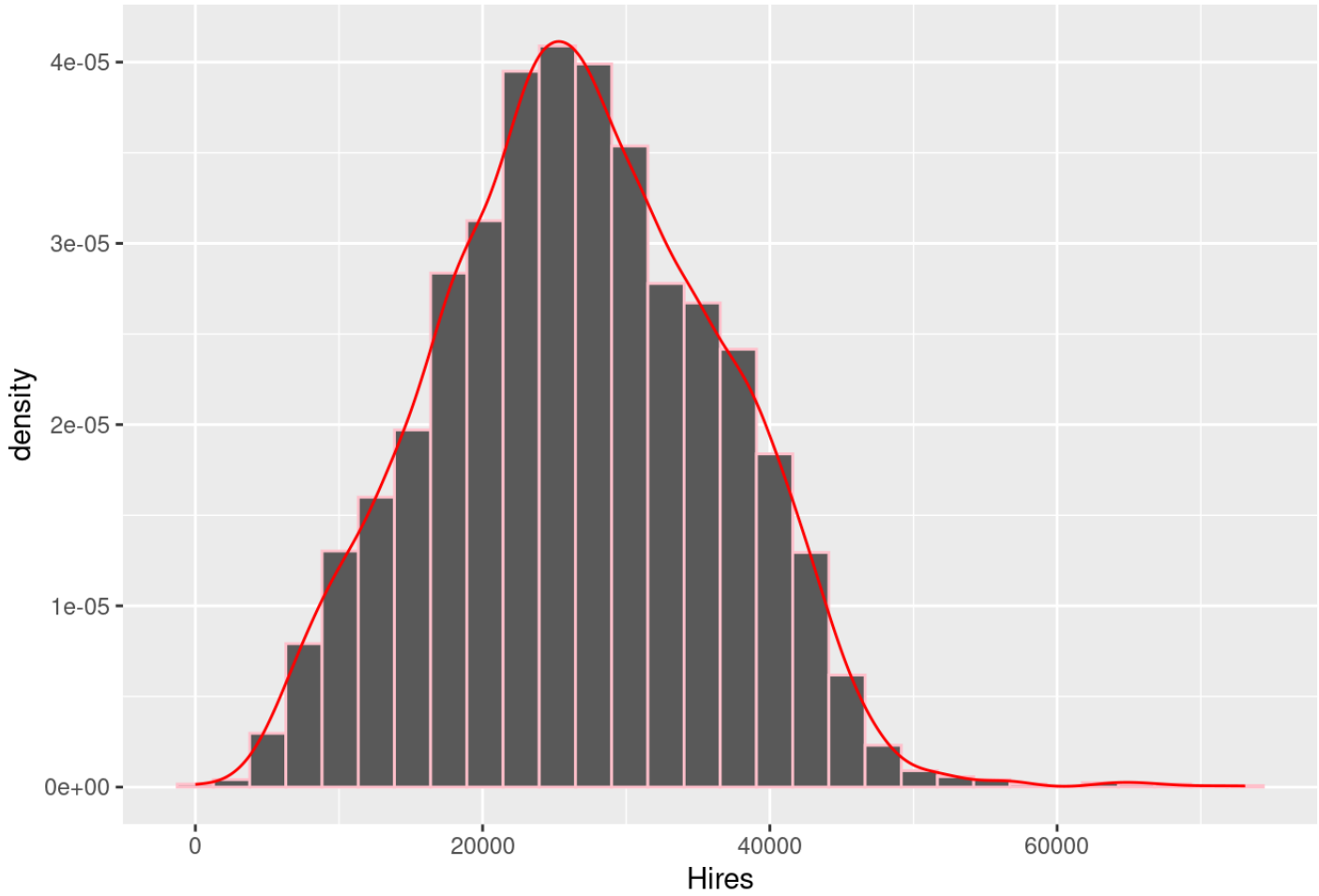


Figure 1: Distribution of bike hires in london from 2010-2023

```
#data shows a normal distribution

#Violin plot + Jitter Plot to show bike hires for each year
hires_plot <- ggplot(data=Covidres_data, aes(x=year, y=Hires)) +
  geom_violin(trim=FALSE,fill= 'pink',alpha=0.8) +
  geom_jitter(size  = 0.7, alpha = 0.3,col="blue") +
  stat_summary(fun.data = mean_sdl,geom="pointrange",col="red",fun.args=list(mult=
1)) +
  labs(title="Number of bike hires in London for each year",subtitle =" Error Bars
show Standard Deviation",x="Year",y="Number of Bike Hires",caption = "Figure 2. Nu
mber of Bikes hired in London for each year")

hires_plot
```

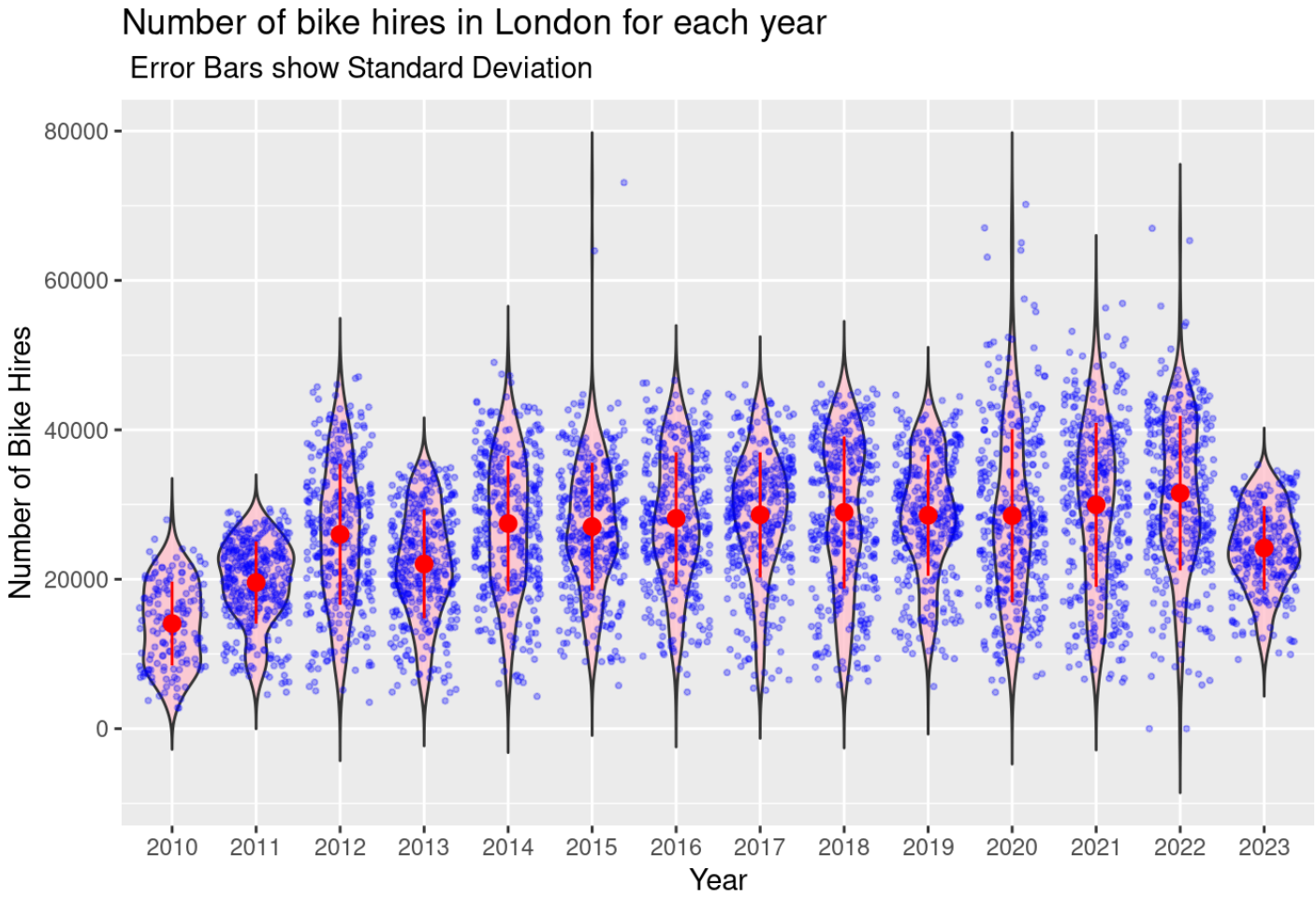


Figure 2. Number of Bikes hired in London for each year

```
#This plot shows the distribution of bike hires from year 2010-2023 along with the mean and standard deviation.
```

We redusce the data size to years 2018 to 2023 to check the effect of COVID restrictions on bike rentals in London.

Our target is to check the effect of covid restrictions on the bike hires so it will be better to cut the data to the years closer to COVID, therefore we just use the data from year 2018 to 2023. We can take 2018 at the reference year for our analysis.

```
# To compare the effects of covid restrictions we filter the table with data from years 2019-2023
covid_hires <- filter( Covidres_data, year %in% c(2018,2019,2020,2021,2022,2023))
covid_hires <- tibble(covid_hires)
str(covid_hires)
```

```
## tibble [2,100 × 8] (S3: tbl_df/tbl/data.frame)
##  $ date               : Date[1:2100], format: "2018-01-01" "2018-01-02" "2018-0
1-03" ...
##  $ Hires              : num [1:2100] 10235 14394 20265 17543 21759 ...
##  $ wfh                : num [1:2100] 0 0 0 0 0 0 0 0 0 0 ...
##  $ rule_of_6_indoors  : num [1:2100] 0 0 0 0 0 0 0 0 0 0 ...
##  $ eat_out_to_help_out: num [1:2100] 0 0 0 0 0 0 0 0 0 0 ...
##  $ day                : Factor w/ 7 levels "Mon","Tue","Wed",..: 1 2 3 4 5 6 7
1 2 3 ...
##  $ month              : Factor w/ 11 levels "Jan","Feb","Mar",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ year               : Factor w/ 14 levels "2010","2011",..: 9 9 9 9 9 9 9 9 9
9 ...
```

Regression analysis to check the effects of the restrictions on hires.

1.Perform t test.

```
# t-test using the data from years 2018 - 2023 shows mean bike hires of 28810, t(2
099) = 12.844, p<0.001
t.test(covid_hires$Hires, mu = 26000)
```

```
##
##  One Sample t-test
##
## data:  covid_hires$Hires
## t = 12.844, df = 2099, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 26000
## 95 percent confidence interval:
##  28380.98 29239.12
## sample estimates:
## mean of x
##  28810.05
```

2.Perform a two sample t-test to see if there is a significant difference in the average bike hires on the days with covid restrictions compared to those with no restrictions.

```
# t-test shows that days with restriction of work from home significantly have les
s bike hires (27974) than those without restriction of work from home  (29719, t(2
091) = 4.02, p<0.001), with a difference of 1745 bikes.
t.test(Hires ~ wfh > 0,data = covid_hires)
```

```
##
##  Welch Two Sample t-test
##
## data:  Hires by wfh > 0
## t = 4.0214, df = 2091.5, p-value = 5.99e-05
## alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
## 95 percent confidence interval:
##   893.9299 2595.6710
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            29719.01            27974.21
```

```
# t-test shows that days with restriction of rule of 6 indoors significantly have
more bike hires (35731) than those without restriction of rule of 6 indoors (2847
8, t(104.77) = -7.153, p<0.001), with a difference of 7253 bikes.
t.test(Hires ~ rule_of_6_indoors > 0,data = covid_hires)
```

```
##
##  Welch Two Sample t-test
##
## data:  Hires by rule_of_6_indoors > 0
## t = -7.1534, df = 104.77, p-value = 1.193e-10
## alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
## 95 percent confidence interval:
##  -9263.179 -5242.373
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            28478.49            35731.27
```

```
# t-test shows that days with restriction of eat out to help out significantly hav
e more bike hires (36418) than those without restriction  (28707, t(28.527) = -5.7
683, p<0.001), with a difference of 7711 bikes.
t.test(Hires ~ eat_out_to_help_out > 0,data = covid_hires)
```

```
##
##  Welch Two Sample t-test
##
## data:  Hires by eat_out_to_help_out > 0
## t = -5.7683, df = 28.527, p-value = 3.197e-06
## alternative hypothesis: true difference in means between group FALSE and group
TRUE is not equal to 0
## 95 percent confidence interval:
##  -10447.271  -4975.118
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            28707.23            36418.43
```

3.Plot the value ratios separately for each year. This suggests that there were some significant changes, and the years of biggest change seem to be the years where restrictions have been enforced, so if we want to better understand the underlying cause we may want to examine how the three restrictions have impacted number of bikes hired in London.

```
m.hire.by.year <- lm(Hires~wfh + rule_of_6_indoors + eat_out_to_help_out + year, d
ata=covid_hires)
summary(m.hire.by.year)
```

```
##
## Call:
## lm(formula = Hires ~ wfh + rule_of_6_indoors + eat_out_to_help_out +
##     year, data = covid_hires)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31523  -6274   -126   6731  44975
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          28952.2      499.2  57.997  < 2e-16 ***
## wfh                  -5836.0      780.0  -7.482 1.07e-13 ***
## rule_of_6_indoors     9548.7     1085.1   8.800  < 2e-16 ***
## eat_out_to_help_out   5387.9     1950.0   2.763 0.005778 **
## year2019              -390.6      706.0  -0.553 0.580091
## year2020              2078.3      896.3   2.319 0.020502 *
## year2021              2842.9      840.8   3.381 0.000735 ***
## year2022              8406.8     1052.0   7.991 2.19e-15 ***
## year2023              1067.9     1091.2   0.979 0.327860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9537 on 2091 degrees of freedom
## Multiple R-squared:  0.09862,    Adjusted R-squared:  0.09518
## F-statistic:  28.6 on 8 and 2091 DF,  p-value: < 2.2e-16
```

```
# The intercept, wfh, rule_of_6_indoors, eat_out_to_help_out, and years 2020, 202
1, and 2022 (p-value <0.05) are statistically significant (indicated by asterisk
s).The year 2019 and 2023 (p-value > 0.01), thus are not statistically significan
t.

(  m.hire.by.year.emm <- emmeans(m.hire.by.year, ~year)  )
```

```
##  year emmean   SE   df lower.CL upper.CL
##  2018  33502 1317 2091    30919    36086
##  2019  33112 1317 2091    30528    35696
##  2020  35581 1051 2091    33519    37643
##  2021  36345 1138 2091    34113    38578
##  2022  41909 1237 2091    39484    44334
##  2023  34570 1270 2091    32079    37061
##
## Results are averaged over the levels of: wfh, rule_of_6_indoors, eat_out_to_hel
p_out
## Confidence level used: 0.95
```

```
(  plot.m.hire.by.year.emm <- ggplot(summary(m.hire.by.year.emm), aes(x=year, y=em
mean, ymin=lower.CL, ymax=upper.CL)) + geom_point(col="black") + geom_linerange(co
l="black") + labs(caption = "Figure 3:Estimated Means of Bike Hires by Year",title
= "Estimated Means of Bike Hires by Year",x="year", y="Number of Bike Hires", subt
itle="Error bars are 95% CIs")   )
```

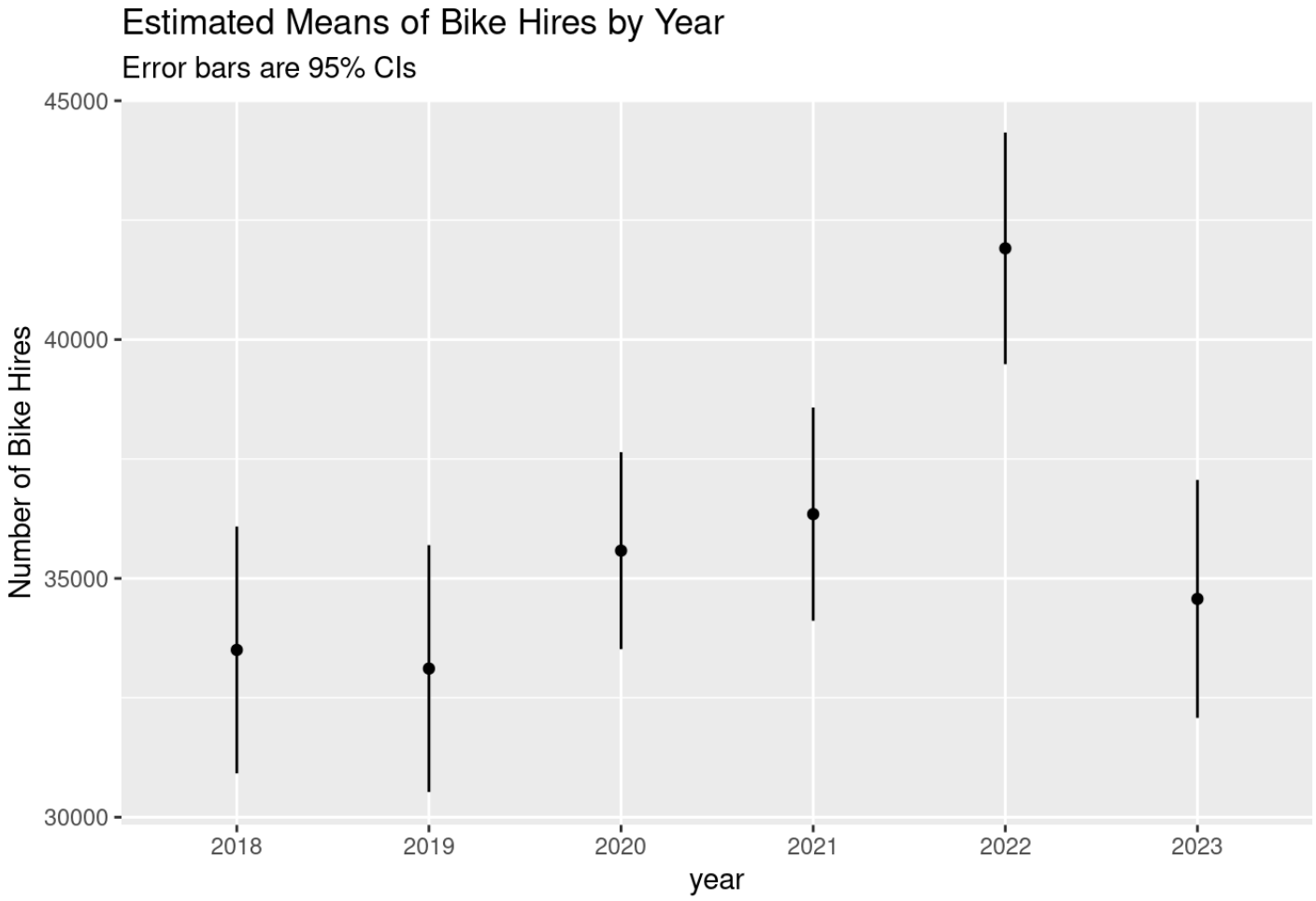


Figure 3:Estimated Means of Bike Hires by Year

```
#The estimated intercept is 28952.2, which represents the predicted value of Hires in year 2018. This intercept is significantly different from zero (p < 2e-16), suggesting a significant intercept term.the model is significant as F-statistic (p-value <0.001), suggesting that the year has a statistically significant association with the number of bike hires.
```

4.Plot the value ratios separately for each month. This suggests that there were some significant changes, and the months of biggest change seem to be the months where restrictions have been enforced, so if we want to better understand the underlying cause we may want to examine how the three restrictions have impacted number of bikes hired in London.

```
m.cnt.by.month <- lm(Hires~wfh + rule_of_6_indoors + eat_out_to_help_out + month, data=covid_hires)

(  m.cnt.by.month.emm <- emmeans(m.cnt.by.month, ~month)  )
```

```
##  month emmean   SE   df lower.CL upper.CL
##  Jan    22285 1064 1900    20199    24370
##  Feb    24435 1078 1900    22320    26549
##  Mar    25784 1064 1900    23697    27871
##  Apr    29778 1070 1900    27679    31878
##  Jun    39152 1014 1900    37162    41141
##  Jul    38904 1030 1900    36885    40923
##  Aug    35905  905 1900    34130    37680
##  Sep    34706 1036 1900    32674    36739
##  Oct    31524 1058 1900    29448    33600
##  Nov    27764 1099 1900    25609    29919
##  Dec    20843 1092 1900    18702    22983
##
## Results are averaged over the levels of: wfh, rule_of_6_indoors, eat_out_to_hel
p_out
## Confidence level used: 0.95
```

```
(  plot.m.cnt.by.month.emm <- ggplot(summary(m.cnt.by.month.emm), aes(x=month, y=e
mmean, ymin=lower.CL, ymax=upper.CL)) + geom_point(col="magenta") + geom_linerang
e(col="magenta") + labs(caption = "Figure 4: Estimated Means of Bike Hires by Mont
h ",title = "Estimated Means of Bike Hires by Month",x="Month", y="Number of Bike
Hires", subtitle="Error bars are 95% CIs")   )
```

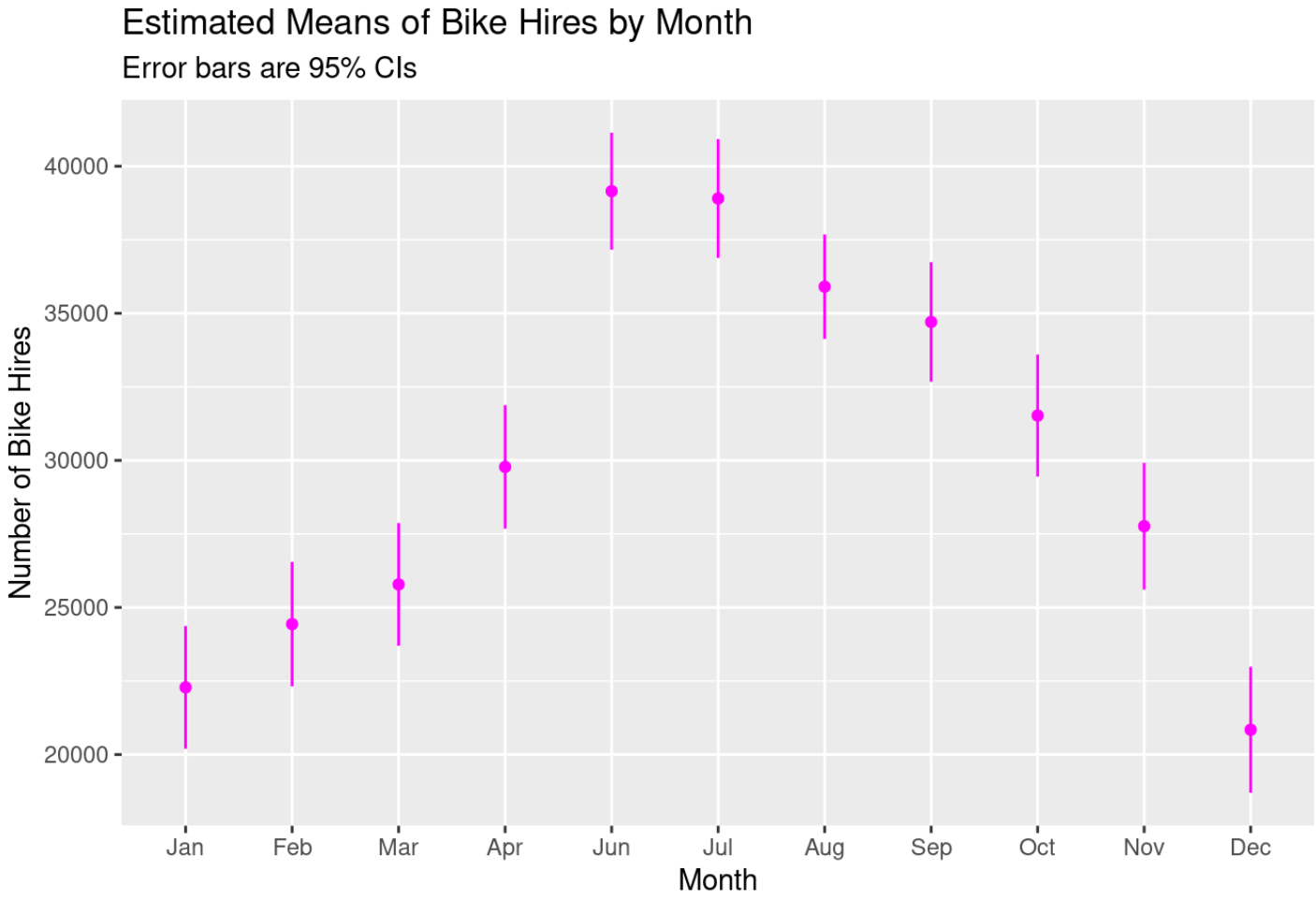


Figure 4: Estimated Means of Bike Hires by Month

```
# Interpreting the individual coefficients
summary(m.cnt.by.month)
```

```
##
## Call:
## lm(formula = Hires ~ wfh + rule_of_6_indoors + eat_out_to_help_out +
##     month, data = covid_hires)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31525  -4993    634   4934  34248
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          21444.0      588.6  36.429  < 2e-16 ***
## wfh                  -2340.7      364.3  -6.424 1.67e-10 ***
## rule_of_6_indoors     2668.2      918.4   2.905  0.00371 **
## eat_out_to_help_out   1354.0     1571.8   0.861  0.38913
## monthFeb              2149.9      811.3   2.650  0.00812 **
## monthMar              3499.4      792.2   4.418 1.05e-05 ***
## monthApr              7493.8      800.5   9.362  < 2e-16 ***
## monthJun             16867.2      813.1  20.744  < 2e-16 ***
## monthJul             16619.2      796.7  20.860  < 2e-16 ***
## monthAug             13620.5      826.9  16.473  < 2e-16 ***
## monthSep             12421.8      804.7  15.436  < 2e-16 ***
## monthOct              9239.2      837.1  11.037  < 2e-16 ***
## monthNov              5478.9      838.5   6.534 8.21e-11 ***
## monthDec             -1442.2      828.8  -1.740  0.08201 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7634 on 1900 degrees of freedom
##   (186 observations deleted due to missingness)
## Multiple R-squared:  0.4187, Adjusted R-squared:  0.4147
## F-statistic: 105.3 on 13 and 1900 DF,  p-value: < 2.2e-16
```

5.Pearson Correlation Matrix

All p-values are very small (<0.001), indicating that the correlations are statistically significant.

```
rcorr(as.matrix(select(covid_hires, Hires, wfh, rule_of_6_indoors, eat_out_to_help
_out)))
```

```
##                     Hires   wfh rule_of_6_indoors eat_out_to_help_out
## Hires                1.00 -0.09              0.15                0.09
## wfh                 -0.09  1.00              0.17               -0.12
## rule_of_6_indoors    0.15  0.17              1.00               -0.03
## eat_out_to_help_out  0.09 -0.12             -0.03                1.00
##
## n= 2100
##
##
## P
##                     Hires  wfh    rule_of_6_indoors eat_out_to_help_out
## Hires                      0.0000 0.0000            0.0000
## wfh                 0.0000        0.0000            0.0000
## rule_of_6_indoors   0.0000 0.0000                   0.2438
## eat_out_to_help_out 0.0000 0.0000 0.2438
```

Correlation Matrix shows: The correlation coefficient is -0.09, is indicating a weak negative correlation between the number of hires and the prevalence of working from home.The correlation coefficient is 0.15, suggesting a weak positive correlation between the number of hires and adherence to the rule_of_6_indoors.The correlation coefficient is -0.12, suggesting a weak negative correlation between working from home and participation in the eat_out_to_help_out scheme.The correlation coefficient is -0.03, indicating a very weak negative correlation between adherence to the rule_of_6_indoors and participation in the eat_out_to_help_out scheme.

6. Linear regression models to test the relation of restrictions with bike hires in London.

```
#model1
m.hire.by.wfh <- lm(Hires ~ wfh, data=covid_hires)
m.hire.by.wfh.emm <- emmeans(m.hire.by.wfh, ~wfh)
summary(m.hire.by.wfh.emm)
```

```
##  wfh emmean  SE   df lower.CL upper.CL
##    0  29719 315 2098    29101    30337
##    1  27974 302 2098    27382    28567
##
## Confidence level used: 0.95
```

```
#The mean bike hires with the restriction of work from home is 27974, 95% CI[27382
--28567] bikes. The mean bike  hires without the restriction of work from home is
29719 bikes, 95% CI[29101--30337] minutes.
(  m.hire.by.wfh.contrast <- confint(pairs(m.hire.by.wfh.emm))  )
```

```
##  contrast    estimate  SE   df lower.CL upper.CL
##  wfh0 - wfh1     1745 436 2098      889     2601
##
## Confidence level used: 0.95
```

 # The mean difference between the number of bike hired without restriction of work from home to with restriction is 1745bikes, 95% CI[889--2601].

```
#model2
m.hires.by.ruleof6 <- lm(Hires ~ rule_of_6_indoors, data=covid_hires)
m.hires.by.ruleof6.emm <- emmeans(m.hires.by.ruleof6, ~rule_of_6_indoors)
summary(m.hires.by.ruleof6.emm)
```

```
##  rule_of_6_indoors emmean   SE   df lower.CL upper.CL
##                  0  28478  221 2098    28044    28913
##                  1  35731 1012 2098    33747    37715
##
## Confidence level used: 0.95
```

#The mean bike hires with the restriction of rule of 6 indoors is 35731, 95% CI[33747--37715] bikes. The mean bike hires without the restriction of rule of 6 indoors is 28478 bikes, 95% CI[28044--28913] minutes.

```
(  m.hire.by.rule6.contrast <- confint(pairs(m.hires.by.ruleof6.emm))  )
```

```
##  contrast                                estimate   SE   df lower.CL upper.CL
##  rule_of_6_indoors0 - rule_of_6_indoors1    -7253 1036 2098    -9284    -5222
##
## Confidence level used: 0.95
```

 # The mean difference between the number of bike hired without restriction of rule of 6 indoors to with restriction is -7253 bikes, 95% CI[-9284 -- -5222].

```
#model3
m.hires.by.eatout <- lm(Hires ~ eat_out_to_help_out, data=covid_hires)
m.hires.by.eatout.emm <- emmeans(m.hires.by.eatout, ~eat_out_to_help_out)
summary(m.hires.by.eatout.emm)
```

```
##  eat_out_to_help_out emmean   SE   df lower.CL upper.CL
##                    0  28707  219 2098    28277    29138
##                    1  36418 1888 2098    32716    40121
##
## Confidence level used: 0.95
```

#The mean bike hires with the restriction of eat out to help out is 35731, 95% CI[33747--37715] bikes. The mean bike hires without the restriction of eat_out_to_help_out is 28478 bikes, 95% CI[28044--28913] minutes.

```
(  m.hires.by.eatout.contrast <- confint(pairs(m.hires.by.eatout.emm))  )
```

```
##  contrast                                        estimate   SE   df lower.CL upper.
CL
##  eat_out_to_help_out0 - eat_out_to_help_out1    -7711 1901 2098   -11438    -39
84
##
## Confidence level used: 0.95
```

```
# The mean difference between the number of bike hired without restriction of eat_
out_to_help_out to with restriction is -7253 bikes, 95% CI[-9284 -- -5222].
```

```
#model4 (Interaction between the restrictions)
m.hires.by.intr <- lm(Hires ~ eat_out_to_help_out*wfh*rule_of_6_indoors , data=cov
id_hires)
summary(m.hires.by.intr)
```

```
##
## Call:
## lm(formula = Hires ~ eat_out_to_help_out * wfh * rule_of_6_indoors,
##     data = covid_hires)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27351  -6247   -371   6871  42819
##
## Coefficients: (3 not defined because of singularities)
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               29419.1      315.2  93.342  < 2e-16
***
## eat_out_to_help_out                        6999.3     1881.7   3.720 0.000205
***
## wfh                                       -2068.5      441.7  -4.683 3.01e-06
***
## rule_of_6_indoors                         13217.4     3484.8   3.793 0.000153
***
## eat_out_to_help_out:wfh                       NA         NA      NA       NA
## eat_out_to_help_out:rule_of_6_indoors         NA         NA      NA       NA
## wfh:rule_of_6_indoors                     -5464.5     3651.7  -1.496 0.134691
## eat_out_to_help_out:wfh:rule_of_6_indoors      NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9816 on 2095 degrees of freedom
## Multiple R-squared:  0.04331,    Adjusted R-squared:  0.04148
## F-statistic: 23.71 on 4 and 2095 DF,  p-value: < 2.2e-16
```

```
cbind(coef(m.hires.by.intr), confint(m.hires.by.intr))
```

```
##                                                          2.5 %   97.5 %
## (Intercept)                                  29419.087  28800.996 30037.18
## eat_out_to_help_out                           6999.342   3309.239 10689.44
## wfh                                          -2068.517  -2934.774 -1202.26
## rule_of_6_indoors                            13217.413   6383.387 20051.44
## eat_out_to_help_out:wfh                             NA         NA       NA
## eat_out_to_help_out:rule_of_6_indoors               NA         NA       NA
## wfh:rule_of_6_indoors                        -5464.460 -12625.700  1696.78
## eat_out_to_help_out:wfh:rule_of_6_indoors           NA         NA       NA
```

```
# This shows that the interaction between the restrictions is statistically not si
gnificant.
```

```
#model5
m.hires.by.all <- lm(Hires ~ eat_out_to_help_out + rule_of_6_indoors + wfh, data=c
ovid_hires)
summary(m.hires.by.all)
```

```
##
## Call:
## lm(formula = Hires ~ eat_out_to_help_out + rule_of_6_indoors +
##     wfh, data = covid_hires)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27311  -6271   -319   6850  42859
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          29459.8      314.1  93.793  < 2e-16 ***
## eat_out_to_help_out   6958.6     1882.0   3.697 0.000223 ***
## rule_of_6_indoors     8240.9     1041.7   7.911 4.08e-15 ***
## wfh                  -2148.5      438.6  -4.898 1.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9819 on 2096 degrees of freedom
## Multiple R-squared:  0.04229,    Adjusted R-squared:  0.04092
## F-statistic: 30.85 on 3 and 2096 DF,  p-value: < 2.2e-16
```

```
cbind(coef(m.hires.by.all), confint(m.hires.by.all))
```

```
##                                  2.5 %    97.5 %
## (Intercept)         29459.794 28843.828 30075.760
## eat_out_to_help_out  6958.634  3267.828 10649.441
## rule_of_6_indoors    8240.912  6198.126 10283.699
## wfh                 -2148.475 -3008.625 -1288.326
```

```
vif(m.hires.by.all)
```

```
## eat_out_to_help_out   rule_of_6_indoors                 wfh
##            1.014936            1.031011            1.045711
```

```
# VIF around 1 suggests a low level of multicollinearity, indicating that this var
iable does not have strong correlations with the other predictors and thus this is
a good model for our analysis.
```

Intercept (29459.8) which represents the expected number of bike hires without any COVID restrictions. For eat out to help out rule t- value is 3.697 (p-value <0.001) is highly significant, indicates the effect of increasing the eat out to help out rule variable by one unit will increase estimated number of bike hires by 6959 units. For Rule of 6 indoors rule t- value is 7.911 (p-value <0.001) is highly significant, indicates the effect of increasing the Rule of 6 indoors rule variable by one unit will increase estimated number of bike hires by 8241 units. For Work from home t- value is -4.898 (p-value <0.001) is highly significant indicates the effect of increasing the work from home rule variable by one unit will decrease estimated number of bike hires by 2148 units.

```
# We use ANOVA to perform a direct model comparison which is often the best approa
ch to compare nested models: a simpler model with a more complex version to find t
he model which best suits our analysis.

anova(m.hires.by.intr, m.hire.by.wfh, m.hires.by.ruleof6, m.hires.by.eatout, m.hir
es.by.all)
```

```
## Analysis of Variance Table
##
## Model 1: Hires ~ eat_out_to_help_out * wfh * rule_of_6_indoors
## Model 2: Hires ~ wfh
## Model 3: Hires ~ rule_of_6_indoors
## Model 4: Hires ~ eat_out_to_help_out
## Model 5: Hires ~ eat_out_to_help_out + rule_of_6_indoors + wfh
##   Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
## 1   2095 2.0187e+11
## 2   2098 2.0941e+11 -3 -7542895398 26.094 < 2.2e-16 ***
## 3   2098 2.0618e+11  0  3223548440
## 4   2098 2.0936e+11  0 -3176263156
## 5   2096 2.0208e+11  2  7279838541 37.776 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# A higher F-statistic suggests that the model is providing a better fit to the da
ta. The associated p-value helps you determine whether the observed F-statistic is
statistically significant. From our anova table we can see that Model 5 : Hires ~
eat_out_to_help_out + rule_of_6_indoors + wfh has the highest F Statistic (37.776)
(p-value <0.001) which makes it a good fit for our data.

# we contrast the impact of the longest imposed covid restriction and pre and post
covid bike hires.
m.hires.con <- aov(Hires ~ year+wfh, data=covid_hires)
m.hires.con.emm <- emmeans(m.hires.con, ~year+wfh)
m.hires.con.emm
```

```
##  year wfh emmean  SE   df lower.CL upper.CL
##  2018   0  28952 509 2093    27954    29951
##  2019   0  28562 509 2093    27563    29560
##  2020   0  31625 701 2093    30249    33000
##  2021   0  32817 674 2093    31495    34139
##  2022   0  36315 901 2093    34549    38081
##  2023   0  28976 948 2093    27117    30835
##  2018   1  24160 901 2093    22394    25927
##  2019   1  23770 901 2093    22004    25536
##  2020   1  26833 571 2093    25713    27952
##  2021   1  28025 591 2093    26867    29183
##  2022   1  31523 509 2093    30525    32521
##  2023   1  24184 589 2093    23030    25338
##
## Confidence level used: 0.95
```

```
(  contrasts <- data.frame(covid.after.effect=c(0,0,1,-1/2,-1/2,0,0,0,1,-1/2,-1/2,
0 ),
                           covid.before.effect=c( -1/2,-1/2,1,0,0,0,-1/2,-1/2,1,0,
0,0))  )
```

```
##    covid.after.effect covid.before.effect
## 1                 0.0                -0.5
## 2                 0.0                -0.5
## 3                 1.0                 1.0
## 4                -0.5                 0.0
## 5                -0.5                 0.0
## 6                 0.0                 0.0
## 7                 0.0                -0.5
## 8                 0.0                -0.5
## 9                 1.0                 1.0
## 10               -0.5                 0.0
## 11               -0.5                 0.0
## 12                0.0                 0.0
```

```
contrast(m.hires.con.emm, contrasts)
```

```
##  contrast            estimate   SE   df t.ratio p.value
##  covid.after.effect     -5882 1265 2093  -4.651  <.0001
##  covid.before.effect     5736 1577 2093   3.638  0.0003
```

```
confint(contrast(m.hires.con.emm, contrasts))
```

```
##  contrast            estimate   SE   df lower.CL upper.CL
##  covid.after.effect     -5882 1265 2093    -8362    -3402
##  covid.before.effect     5736 1577 2093     2644     8827
##
## Confidence level used: 0.95
```

There is a decrease in bike hires from 2020 to 2022 (immediately before and after the work from home restriction) by 6000 bike less 95% CI[-8362 – -3402]. Whereas the number of bike hires before the work from hone restriction were 5736 more 95% CI [2644 – 8827] This shows that the restriction of work from home has negetively impacted the number of bikes rented and we still haven't reach the pre covid levels.

7.Visualsation of the three restrictions to check the effect on bike hires yearly, monthly and weekly.

```
covid_hires <- covid_hires %>%
  mutate(
    Covid = case_when(
      wfh == 1 ~ "WFH",
      rule_of_6_indoors == 1 ~ "Rule of 6 Indoors",
      eat_out_to_help_out == 1 ~ "Eat Out to Help Out",
      TRUE ~ "No COVID"
    )
  )
#year
m.Hires.by.year.and.Covidres <- glm(Hires~year * Covid, data=covid_hires)
summary(m.Hires.by.year.and.Covidres)
```

```
##
## Call:
## glm(formula = Hires ~ year * Covid, data = covid_hires)
##
## Coefficients: (14 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     28532.6     2222.9  12.836  < 2e-16 ***
## year2019                         -390.6      708.1  -0.552   0.5812
## year2020                         7885.8     1293.5   6.096 1.29e-09 ***
## year2021                         6130.5      932.2   6.576 6.07e-11 ***
## year2022                        11166.8     1240.7   9.000  < 2e-16 ***
## year2023                         3828.0     1274.3   3.004   0.0027 **
## CovidNo COVID                     419.5     2165.8   0.194   0.8464
## CovidRule of 6 Indoors           6218.1     3834.9   1.621   0.1051
## CovidWFH                        -8176.5     1911.2  -4.278 1.97e-05 ***
## year2019:CovidNo COVID               NA         NA      NA       NA
## year2020:CovidNo COVID          -11275.1     1554.7  -7.252 5.75e-13 ***
## year2021:CovidNo COVID               NA         NA      NA       NA
## year2022:CovidNo COVID               NA         NA      NA       NA
## year2023:CovidNo COVID               NA         NA      NA       NA
## year2019:CovidRule of 6 Indoors      NA         NA      NA       NA
## year2020:CovidRule of 6 Indoors      NA         NA      NA       NA
## year2021:CovidRule of 6 Indoors      NA         NA      NA       NA
## year2022:CovidRule of 6 Indoors      NA         NA      NA       NA
## year2023:CovidRule of 6 Indoors      NA         NA      NA       NA
## year2019:CovidWFH                    NA         NA      NA       NA
## year2020:CovidWFH                    NA         NA      NA       NA
## year2021:CovidWFH                    NA         NA      NA       NA
## year2022:CovidWFH                    NA         NA      NA       NA
## year2023:CovidWFH                    NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 91506442)
##
##     Null deviance: 2.1100e+11  on 2099  degrees of freedom
## Residual deviance: 1.9125e+11  on 2090  degrees of freedom
## AIC: 44469
##
## Number of Fisher Scoring iterations: 2
```

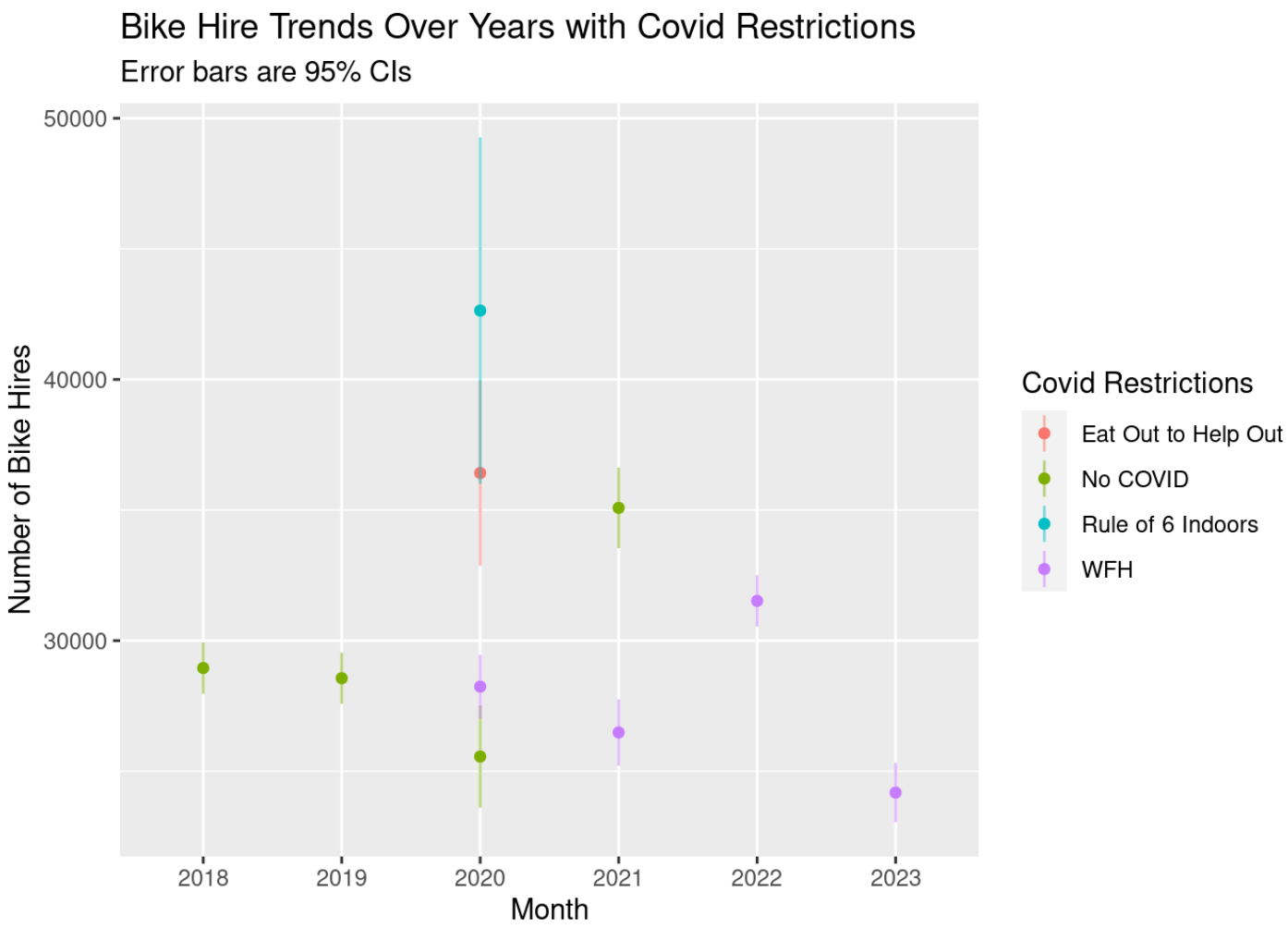
```
( m.Hires.by.year.and.Covidres.emm <- emmeans(m.Hires.by.year.and.Covidres, ~year
+ Covid)  )
```

```
##  year Covid                emmean   SE   df lower.CL upper.CL
##  2018 Eat Out to Help Out nonEst   NA   NA       NA       NA
##  2019 Eat Out to Help Out nonEst   NA   NA       NA       NA
##  2020 Eat Out to Help Out  36418 1808 2090    32873    39964
##  2021 Eat Out to Help Out nonEst   NA   NA       NA       NA
##  2022 Eat Out to Help Out nonEst   NA   NA       NA       NA
##  2023 Eat Out to Help Out nonEst   NA   NA       NA       NA
##  2018 No COVID             28952  501 2090    27970    29934
##  2019 No COVID             28562  501 2090    27580    29543
##  2020 No COVID             25563  997 2090    23607    27519
##  2021 No COVID             35083  786 2090    33541    36625
##  2022 No COVID            nonEst   NA   NA       NA       NA
##  2023 No COVID            nonEst   NA   NA       NA       NA
##  2018 Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  2019 Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  2020 Rule of 6 Indoors    42636 3382 2090    36004    49269
##  2021 Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  2022 Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  2023 Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  2018 WFH                 nonEst   NA   NA       NA       NA
##  2019 WFH                 nonEst   NA   NA       NA       NA
##  2020 WFH                  28242  620 2090    27026    29458
##  2021 WFH                  26487  648 2090    25216    27757
##  2022 WFH                  31523  501 2090    30541    32505
##  2023 WFH                  24184  579 2090    23049    25319
##
## Confidence level used: 0.95
```

```
ggplot(summary(m.Hires.by.year.and.Covidres.emm), aes(x=year, y=emmean, ymin=lowe
r.CL, ymax=upper.CL, col=Covid)) + geom_point() + geom_linerange(alpha=0.5) + lab
s(Caption = "Figure 5:Bike Hire Trends Over Years with Covid Restrictions ",title
= "Bike Hire Trends Over Years with Covid Restrictions", x="Month", y="Number of B
ike Hires", col="Covid Restrictions", subtitle="Error bars are 95% CIs")
```

```
## Warning: Removed 14 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 14 rows containing missing values (`geom_segment()`).
```




```
## The graph shows how the three restrictions effected the bike hires with differe
nt year.

#months
m.Hires.by.month.and.Covidres <- glm(Hires~month * Covid, data=covid_hires)
summary(m.Hires.by.month.and.Covidres)
```

```
##
## Call:
## glm(formula = Hires ~ month * Covid, data = covid_hires)
##
## Coefficients: (20 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     22140.1     1890.0  11.714  < 2e-16 ***
## monthFeb                         3715.7     1137.1   3.268 0.001103 **
## monthMar                         5701.5     1068.6   5.335 1.07e-07 ***
## monthApr                         8178.2     1043.6   7.836 7.67e-15 ***
## monthJun                        18962.2     1043.6  18.170  < 2e-16 ***
## monthJul                        17354.3     1061.9  16.342  < 2e-16 ***
## monthAug                        14278.3     1238.5  11.528  < 2e-16 ***
## monthSep                         9552.4     1200.3   7.959 2.97e-15 ***
## monthOct                         9025.6     1238.5   7.287 4.63e-13 ***
```

```
## monthNov                          6166.3     1250.9   4.930 8.96e-07 ***
## monthDec                         -1064.5     1148.1  -0.927 0.353931
## CovidNo COVID                     -189.9     2045.9  -0.093 0.926051
## CovidRule of 6 Indoors           10943.9     3304.3   3.312 0.000944 ***
## CovidWFH                         -3543.1     1720.0  -2.060 0.039540 *
## monthFeb:CovidNo COVID           -3119.2     1605.6  -1.943 0.052195 .
## monthMar:CovidNo COVID           -5056.4     1577.1  -3.206 0.001368 **
## monthApr:CovidNo COVID           -1547.1     1629.1  -0.950 0.342403
## monthJun:CovidNo COVID           -4444.5     1629.1  -2.728 0.006426 **
## monthJul:CovidNo COVID            -939.6     1581.8  -0.594 0.552577
## monthAug:CovidNo COVID           -1261.9     1655.9  -0.762 0.446122
## monthSep:CovidNo COVID            4383.6     1615.0   2.714 0.006703 **
## monthOct:CovidNo COVID             646.3     1661.7   0.389 0.697357
## monthNov:CovidNo COVID           -1314.6     1677.0  -0.784 0.433214
## monthDec:CovidNo COVID            -745.0     1640.9  -0.454 0.649862
## monthFeb:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthMar:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthApr:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthJun:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthJul:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthAug:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthSep:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthOct:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthNov:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthDec:CovidRule of 6 Indoors       NA         NA      NA       NA
## monthFeb:CovidWFH                     NA         NA      NA       NA
## monthMar:CovidWFH                     NA         NA      NA       NA
## monthApr:CovidWFH                     NA         NA      NA       NA
## monthJun:CovidWFH                     NA         NA      NA       NA
## monthJul:CovidWFH                     NA         NA      NA       NA
## monthAug:CovidWFH                     NA         NA      NA       NA
## monthSep:CovidWFH                     NA         NA      NA       NA
## monthOct:CovidWFH                     NA         NA      NA       NA
## monthNov:CovidWFH                     NA         NA      NA       NA
## monthDec:CovidWFH                     NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 57064800)
##
##     Null deviance: 1.9048e+11  on 1913  degrees of freedom
## Residual deviance: 1.0785e+11  on 1890  degrees of freedom
##   (186 observations deleted due to missingness)
## AIC: 39641
##
## Number of Fisher Scoring iterations: 2
```

```
( m.Hires.by.month.and.Covidres.emm <- emmeans(m.Hires.by.month.and.Covidres, ~mon
th + Covid)  )
```

```
##  month Covid                emmean   SE   df lower.CL upper.CL
##  Jan   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Feb   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Mar   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Apr   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Jun   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Jul   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Aug   Eat Out to Help Out  36418 1428 1890    33619    39218
##  Sep   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Oct   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Nov   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Dec   Eat Out to Help Out nonEst   NA   NA       NA       NA
##  Jan   No COVID             21950  783 1890    20414    23487
##  Feb   No COVID             22547  819 1890    20940    24154
##  Mar   No COVID             22595  855 1890    20918    24273
##  Apr   No COVID             28581  975 1890    26669    30494
##  Jun   No COVID             36468  975 1890    34555    38381
##  Jul   No COVID             38365  872 1890    36654    40076
##  Aug   No COVID             34967  771 1890    33455    36479
##  Sep   No COVID             35886  744 1890    34426    37346
##  Oct   No COVID             31622  783 1890    30086    33158
##  Nov   No COVID             26802  796 1890    25240    28364
##  Dec   No COVID             20141  872 1890    18430    21851
##  Jan   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Feb   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Mar   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Apr   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Jun   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Jul   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Aug   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Sep   Rule of 6 Indoors    42636 2671 1890    37398    47874
##  Oct   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Nov   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Dec   Rule of 6 Indoors   nonEst   NA   NA       NA       NA
##  Jan   WFH                  18597  783 1890    17061    20133
##  Feb   WFH                  22313  824 1890    20696    23929
##  Mar   WFH                  24299  727 1890    22873    25724
##  Apr   WFH                  26775  690 1890    25423    28128
##  Jun   WFH                  37559  690 1890    36207    38912
##  Jul   WFH                  35951  717 1890    34545    37358
##  Aug   WFH                  32875  959 1890    30994    34757
##  Sep   WFH                  28149  909 1890    26366    29933
##  Oct   WFH                  27623  959 1890    25741    29504
##  Nov   WFH                  24763  975 1890    22851    26676
##  Dec   WFH                  17532  839 1890    15886    19179
##
## Confidence level used: 0.95
```

```
ggplot(summary(m.Hires.by.month.and.Covidres.emm), aes(x=month, y=emmean, ymin=lower.CL, ymax=upper.CL, col=Covid)) + geom_point() + geom_linerange(alpha=0.5) + labs(caption = "Figure 6: Bike Hire Trends Over different Months with Covid Restrictions ",title = "Bike Hire Trends Over different Months with Covid Restrictions",x="Month", y="Number of Bike Hires", col="Covid Restrictions", subtitle="Error bars are 95% CIs")
```

```
## Warning: Removed 20 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 20 rows containing missing values (`geom_segment()`).
```

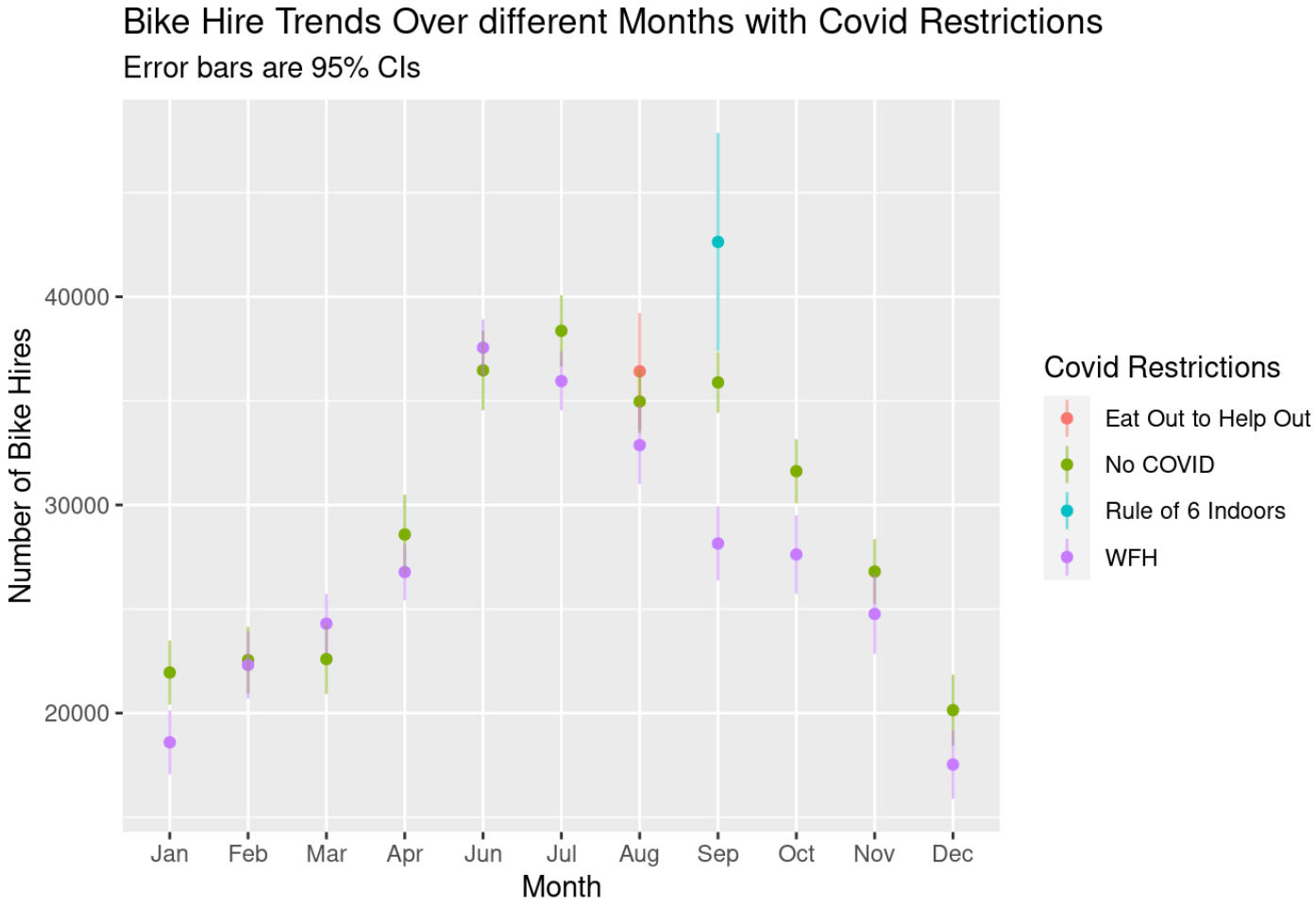


Figure 6: Bike Hire Trends Over different Months with Covid Restrictions

```
# The graph shows how the three restrictions effected the bike hires in different months.

#day
m.Hires.by.day.and.Covidres <- glm(Hires~day * Covid, data=covid_hires)
summary(m.Hires.by.day.and.Covidres)
```

```
##
## Call:
## glm(formula = Hires ~ day * Covid, data = covid_hires)
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   35144.75    4849.86   7.247    6e-13 ***
## dayTue                         2724.00    6858.74   0.397   0.6913
## dayWed                         1482.50    6858.74   0.216   0.8289
## dayThu                          389.00    6858.74   0.057   0.9548
## dayFri                         -914.00    6858.74  -0.133   0.8940
## daySat                         2903.25    6858.74   0.423   0.6721
## daySun                         2331.00    6858.74   0.340   0.7340
## CovidNo COVID                 -5256.41    4918.66  -1.069   0.2853
## CovidRule of 6 Indoors           58.75    8400.21   0.007   0.9944
## CovidWFH                      -9231.96    4912.04  -1.879   0.0603 .
## dayTue:CovidNo COVID           -436.70    6956.74  -0.063   0.9500
## dayWed:CovidNo COVID            926.63    6956.74   0.133   0.8940
## dayThu:CovidNo COVID           1587.75    6956.74   0.228   0.8195
## dayFri:CovidNo COVID           1313.95    6956.74   0.189   0.8502
## daySat:CovidNo COVID          -6884.00    6956.38  -0.990   0.3225
## daySun:CovidNo COVID          -8647.11    6956.38  -1.243   0.2140
## dayTue:CovidRule of 6 Indoors  5664.50   13717.49   0.413   0.6797
## dayWed:CovidRule of 6 Indoors  6669.00   13717.49   0.486   0.6269
## dayThu:CovidRule of 6 Indoors  6440.50   13717.49   0.470   0.6388
## dayFri:CovidRule of 6 Indoors  7816.50   13717.49   0.570   0.5689
## daySat:CovidRule of 6 Indoors 13657.25   13717.49   0.996   0.3196
## daySun:CovidRule of 6 Indoors 10300.50   13717.49   0.751   0.4528
## dayTue:CovidWFH                 -98.31    6946.12  -0.014   0.9887
## dayWed:CovidWFH                 924.91    6946.12   0.133   0.8941
## dayThu:CovidWFH                2357.72    6946.12   0.339   0.7343
## dayFri:CovidWFH                2869.96    6946.12   0.413   0.6795
## daySat:CovidWFH                1278.49    6946.40   0.184   0.8540
## daySun:CovidWFH               -1851.49    6946.68  -0.267   0.7899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 94084731)
##
##     Null deviance: 2.1100e+11  on 2099  degrees of freedom
## Residual deviance: 1.9494e+11  on 2072  degrees of freedom
## AIC: 44545
##
## Number of Fisher Scoring iterations: 2
```

```
( m.Hires.by.day.and.Covidres.emm <- emmeans(m.Hires.by.day.and.Covidres, ~day + C
ovid)  )
```

```
##  day Covid               emmean   SE   df lower.CL upper.CL
##  Mon Eat Out to Help Out  35145 4850 2072    25634    44656
##  Tue Eat Out to Help Out  37869 4850 2072    28358    47380
##  Wed Eat Out to Help Out  36627 4850 2072    27116    46138
##  Thu Eat Out to Help Out  35534 4850 2072    26023    45045
##  Fri Eat Out to Help Out  34231 4850 2072    24720    43742
##  Sat Eat Out to Help Out  38048 4850 2072    28537    47559
##  Sun Eat Out to Help Out  37476 4850 2072    27965    46987
##  Mon No COVID             29888  820 2072    28281    31496
##  Tue No COVID             32176  826 2072    30556    33795
##  Wed No COVID             32297  826 2072    30678    33917
##  Thu No COVID             31865  826 2072    30246    33484
##  Fri No COVID             30288  826 2072    28669    31908
##  Sat No COVID             25908  823 2072    24294    27521
##  Sun No COVID             23572  823 2072    21959    25186
##  Mon Rule of 6 Indoors    35204 6859 2072    21753    48654
##  Tue Rule of 6 Indoors    43592 9700 2072    24570    62614
##  Wed Rule of 6 Indoors    43355 9700 2072    24333    62377
##  Thu Rule of 6 Indoors    42033 9700 2072    23011    61055
##  Fri Rule of 6 Indoors    42106 9700 2072    23084    61128
##  Sat Rule of 6 Indoors    51764 9700 2072    32742    70786
##  Sun Rule of 6 Indoors    47835 9700 2072    28813    66857
##  Mon WFH                  25913  779 2072    24385    27441
##  Tue WFH                  28538  774 2072    27020    30057
##  Wed WFH                  28320  774 2072    26802    29838
##  Thu WFH                  28660  774 2072    27141    30178
##  Fri WFH                  27869  774 2072    26351    29387
##  Sat WFH                  30095  777 2072    28572    31618
##  Sun WFH                  26392  779 2072    24864    27920
##
## Confidence level used: 0.95
```

```
ggplot(summary(m.Hires.by.day.and.Covidres.emm), aes(x=day, y=emmean, ymin=lower.CL, ymax=upper.CL, col=Covid)) + geom_point() + geom_linerange(alpha=0.5) + labs(caption = "Figure 7: Bike Hire Trends Over different days of week with Covid Restrictions",title = "Bike Hire Trends Over different days of week with Covid Restrictions",x="Month", y="Number of Bike Hires", col="Covid Restrictions", subtitle="Error bars are 95% CIs")
```

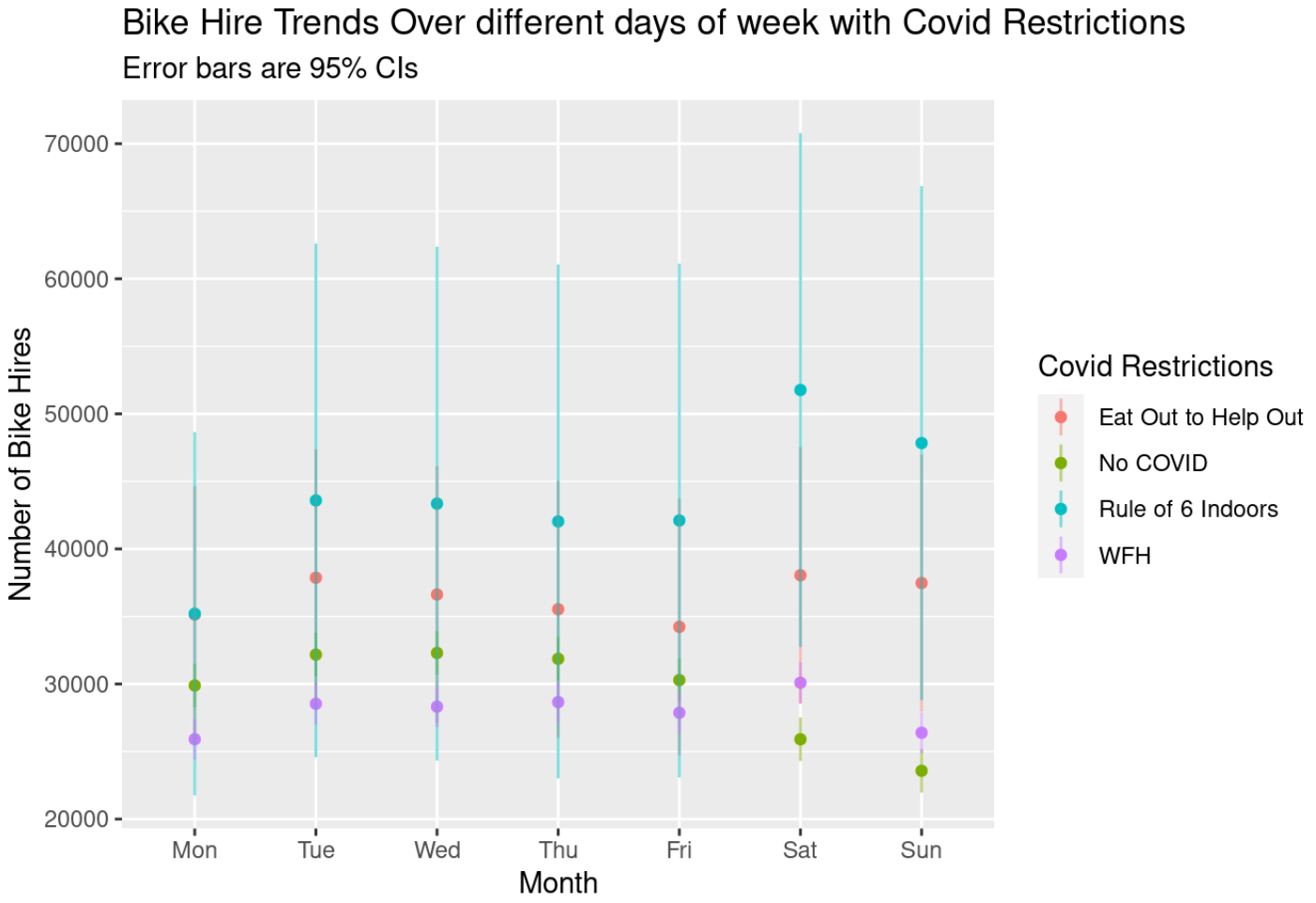


Figure 7: Bike Hire Trends Over different days of week with Covid Restrictions

```
## The graph shows how the three restrictions effected the bike hires with differe
nt days of the week.
```

We can see the from the overall in terms of statistical tests and plots that there is a difference in pattern of bike hires across months with COVID restrictions and without. Showing that work from home had major impact on the number of bike hires which is because it is the longest imposed COVID restriction, showing that for the months from January to June number of bike hires were almost same with and without the restriction, whereas for the months from July to December bike hires were more when there were no restrictions.

Report for London bike hires and covid restrictions

This report analyzes the impact of three COVID-19 restrictions—Work From Home (WFH), Rule of 6 Indoors, and Eat Out to Help Out—on the number of bike hires in London. The data set spans from 2010-07-30 to 2023-09-30, providing daily counts of bike hires along with information on various COVID restrictions.We read the data, check for duplicates, and verify that there are no missing values. The data summary and histogram indicate a normally distributed density without outliers.

We observed that Work From Home is the longest imposed COVID restriction and thus it will have a major impact on our findings about the bike hires in London among the three restrictions that we are considering.

Visualizations include a histogram of overall bike hires, a violin plot showing yearly distribution with jitter points, and a summary of hire distribution for each year.

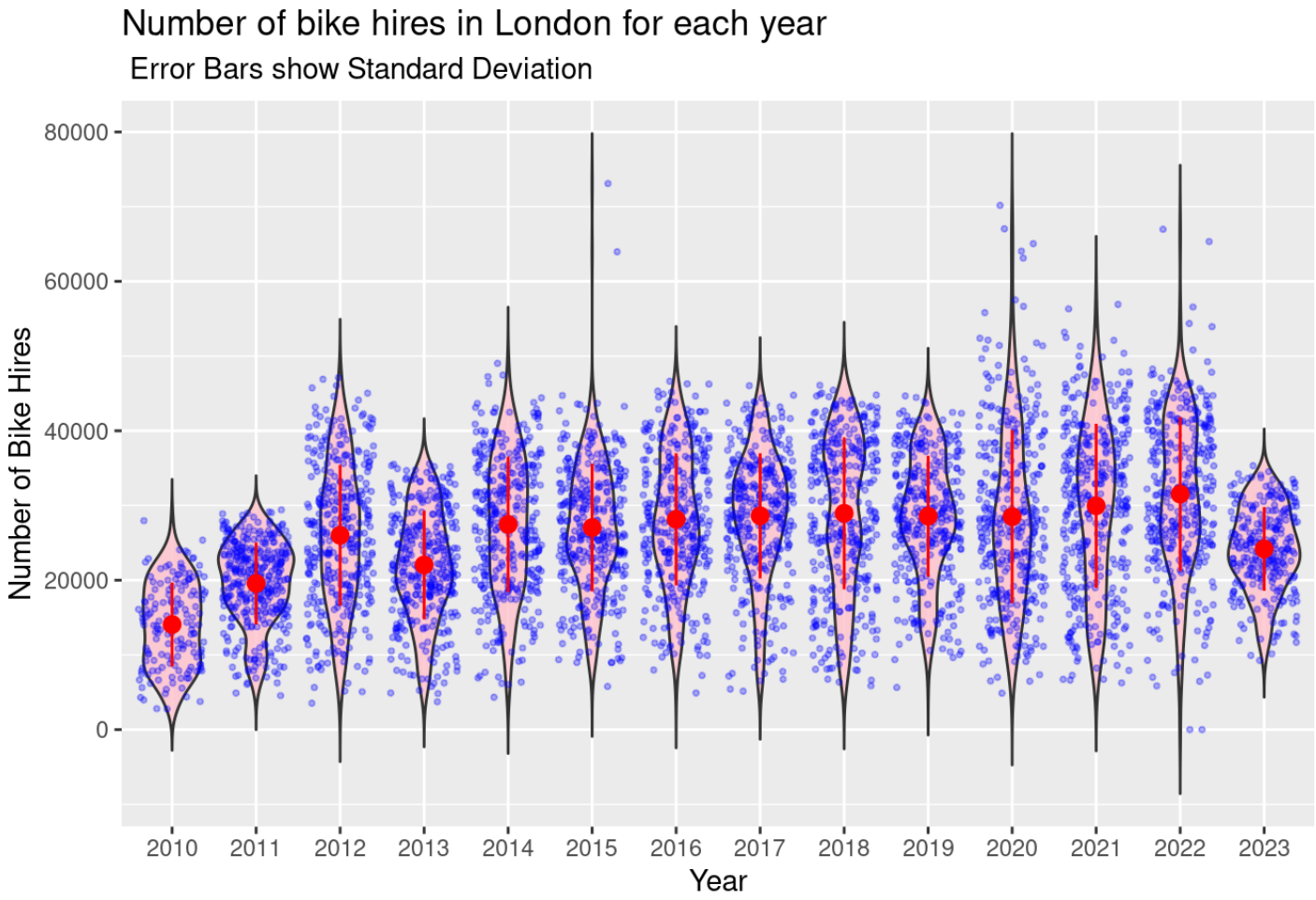
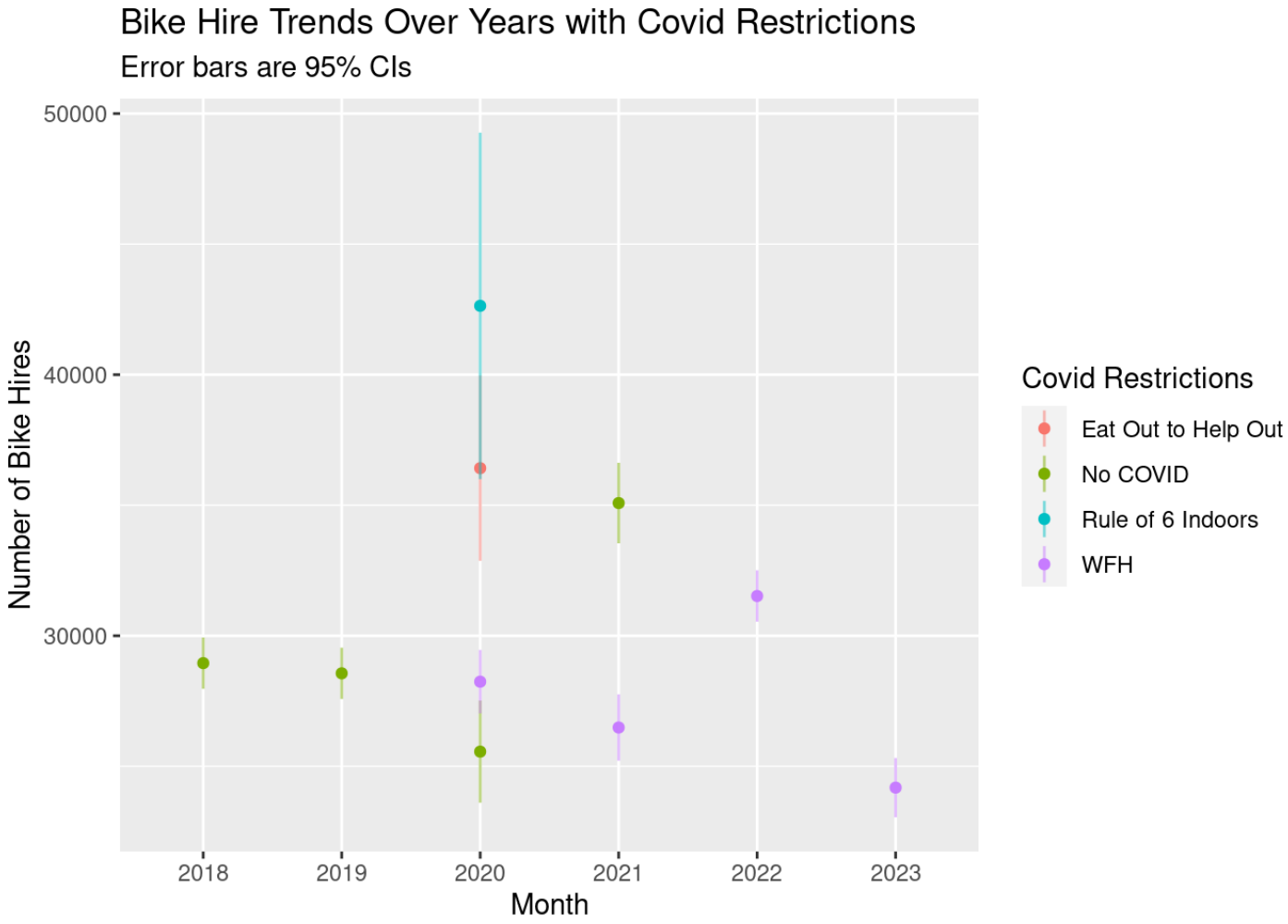


Figure 2. Number of Bikes hired in London for each year

We focus on the years 2018 to 2023 to analyze the effect of COVID restrictions on bike hires. T-tests and linear regression models are performed, showing significant differences in hires with and without specific restrictions.

ANOVA is used to compare nested models, with Model including all the restrictions (Hires ~ eat_out_to_help_out + rule_of_6_indoors + wfh) identified as the best fit with f- measure 37.776 (p-value < 2.2e-16).

The plots below shows the effect of the three COVID restrictions on bike hires yearly from 2018- 2023(data only till September for the year 2023), monthly, and weekly, highlighting differences in trends.



According to the graph the reasn for higher bike hires in year 2020 was due to the covid restrictions imposed.

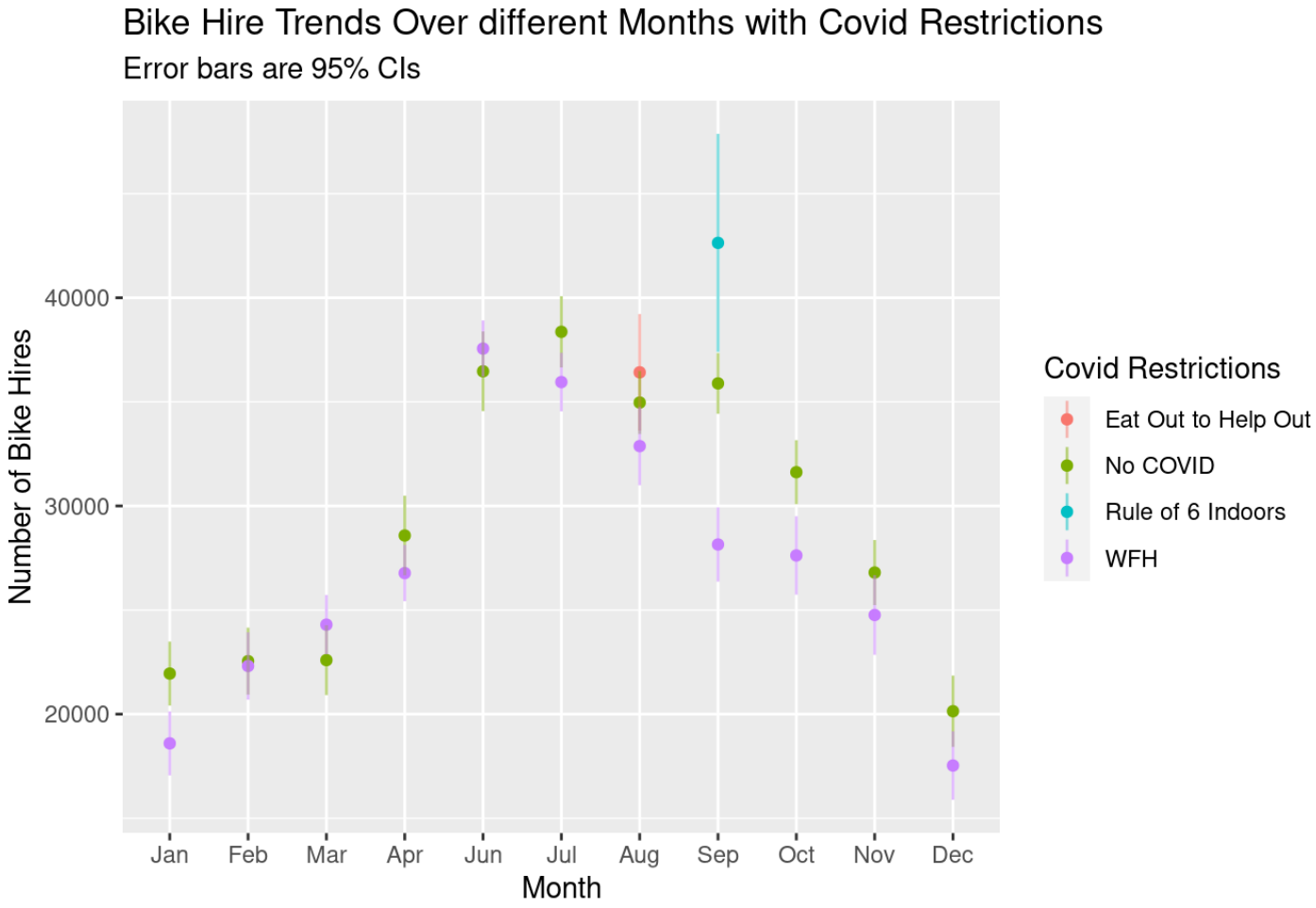


Figure 6: Bike Hire Trends Over different Months with Covid Restrictions

Monthly data shows that work from home has impacted the hires of the bikes in London in a negetive way where as rule of 6 indoors and eat out to help out scheme had impacted in a positive way, increasing the number of bike hires.

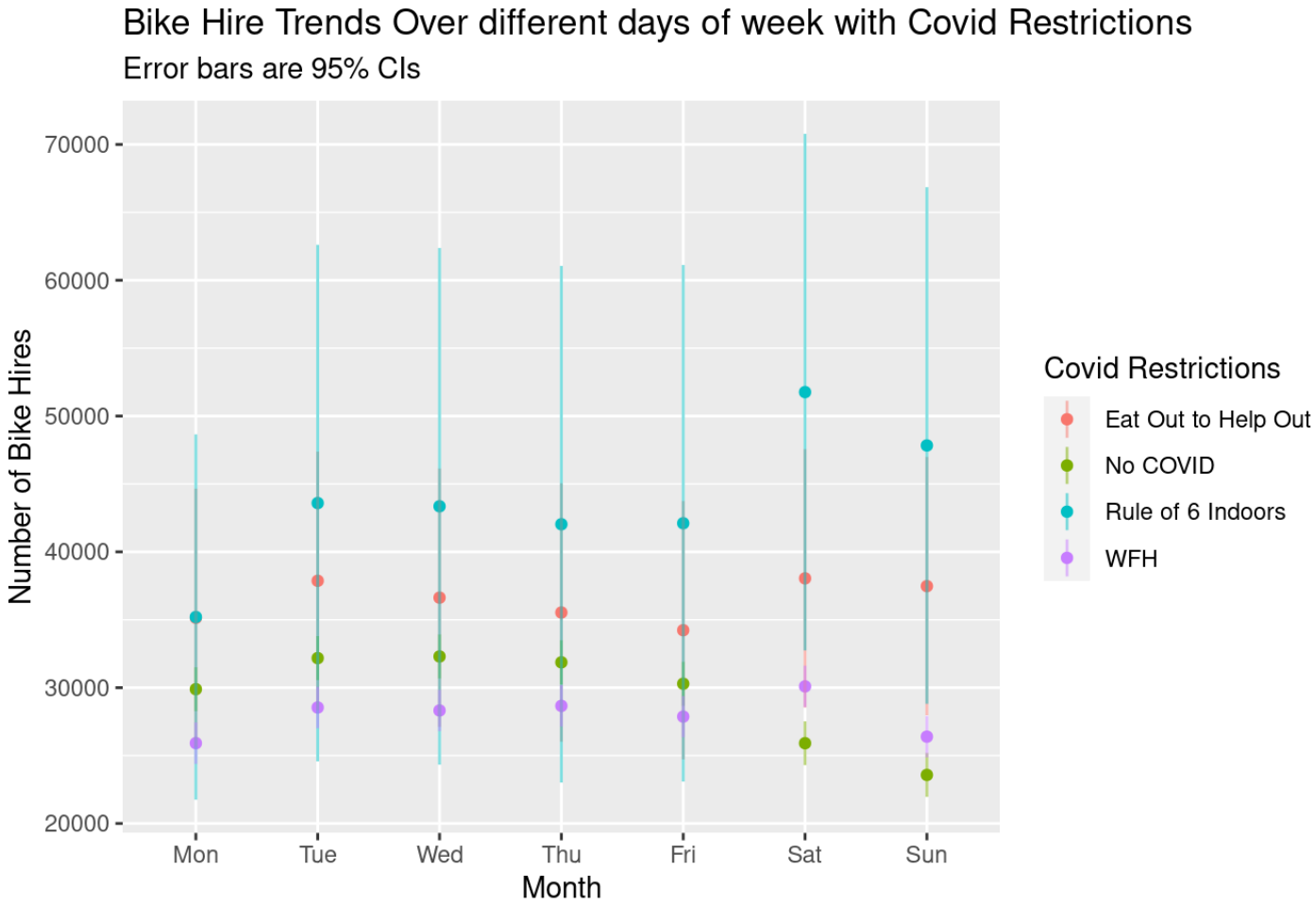


Figure 7: Bike Hire Trends Over different days of week with Covid Restrictions

We can see that on the weekends the number of bikes hired are more for all the three restrictions which could be due to obvious reasons.

On contrasting the Pre Covid and Post Covid time we found out that there is a decrease in bike hires from 2020 to 2022 (immediately before and after the work from home restriction) by 6000 bike less 95% CI[-8362 – -3402]. Whereas the number of bike hires before the work from hone restriction were 5736 more 95% CI [2644 – 8827] This shows that the restriction of work from home has negetively impacted the number of bikes rented and we still haven't reach the pre covid levels.

The analysis reveals significant associations between COVID restrictions and bike hires in London. Work From Home, Rule of 6 Indoors, and Eat Out to Help Out show varying impacts, with detailed insights into monthly and yearly patterns.Work from home reduced commuting-related cycling, while Rule of 6 Indoors and Eat Out to Help Out boosted leisure cycling. Further investigation into specific months and days with the most significant changes in bike hires may provide deeper insights. Consideration of additional factors such as weather conditions could enhance the analysis.

Question 2 – Book Sales

Effect of of reviews on the sales of the book and effect on the sales on the price of books across different gernes.

Data Dictionary

This data provides information on the sales of e-book over a period of many months. where each line of data represents one book.

Variable	Description
sold by	publisher that sold the e-book
publisher.type	publisher type
genre	genre of the book
avg.review	average of the reviews for the book
daily.sales	average number of sales (minus refunds) across all days in the period
total.reviews	tatal reviews for the book
sale.price	average price for which the book sold in the period

Read Data

```
#read data
sales_data <- read_csv("publisher_sales.csv")
```

```
## Rows: 6000 Columns: 7
## ── Column specification ────────────────────────────────────────────────
────────────────────
## Delimiter: ","
## chr (3): sold by, publisher.type, genre
## dbl (4): avg.review, daily.sales, total.reviews, sale.price
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this messag
e.
```

Data Quality Check

```
#check data summary
str(sales_data)
```

```
## spc_tbl_ [6,000 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ sold by       : chr [1:6000] "Amazon Digital Services,  Inc." "HarperCollins
Publishers" "Amazon Digital Services,  Inc." "Amazon Digital Services,  Inc." ...
##  $ publisher.type: chr [1:6000] "indie" "big five" "small/medium" "small/mediu
m" ...
##  $ genre         : chr [1:6000] "adult_fiction" "YA_fiction" "adult_fiction" "Y
A_fiction" ...
##  $ avg.review    : num [1:6000] 4.5 4.64 2.5 4.5 4.98 3.98 4.62 3.5 4.64 4.56
...
##  $ daily.sales   : num [1:6000] 84.4 113.1 70.8 149.4 135.7 ...
##  $ total.reviews : num [1:6000] 151 184 125 225 194 123 130 110 129 119 ...
##  $ sale.price    : num [1:6000] 5.12 6.91 6.27 4.91 7.39 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `sold by` = col_character(),
##   ..   publisher.type = col_character(),
##   ..   genre = col_character(),
##   ..   avg.review = col_double(),
##   ..   daily.sales = col_double(),
##   ..   total.reviews = col_double(),
##   ..   sale.price = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#check for duplicates
sum(duplicated(sales_data))
```

```
## [1] 0
```

```
#check for missing values
sum(is.na(sales_data))
```

```
## [1] 0
```

Data Visualisation

```
#check data summary
str(sales_data)
```

```
## spc_tbl_ [6,000 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ sold by       : chr [1:6000] "Amazon Digital Services,  Inc." "HarperCollins
Publishers" "Amazon Digital Services,  Inc." "Amazon Digital Services,  Inc." ...
##  $ publisher.type: chr [1:6000] "indie" "big five" "small/medium" "small/mediu
m" ...
##  $ genre         : chr [1:6000] "adult_fiction" "YA_fiction" "adult_fiction" "Y
A_fiction" ...
##  $ avg.review    : num [1:6000] 4.5 4.64 2.5 4.5 4.98 3.98 4.62 3.5 4.64 4.56
...
##  $ daily.sales   : num [1:6000] 84.4 113.1 70.8 149.4 135.7 ...
##  $ total.reviews : num [1:6000] 151 184 125 225 194 123 130 110 129 119 ...
##  $ sale.price    : num [1:6000] 5.12 6.91 6.27 4.91 7.39 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `sold by` = col_character(),
##   ..   publisher.type = col_character(),
##   ..   genre = col_character(),
##   ..   avg.review = col_double(),
##   ..   daily.sales = col_double(),
##   ..   total.reviews = col_double(),
##   ..   sale.price = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#check for duplicates
sum(duplicated(sales_data))
```

```
## [1] 0
```

```
#check for missing values
sum(is.na(sales_data))
```

```
## [1] 0
```

```
sum(sales_data$total.reviews ==0)
```

```
## [1] 23
```

```
sum(sales_data$avg.review ==0)
```

```
## [1] 23
```

```
#Most of the distribution density is normally distributed, without any high values
that could potentially be designated as outliers.
#Simple histogram showing the overall distribution and checking for outliers.
p.sales.distribution <- ggplot(sales_data,aes(x= daily.sales,..density..)) +
  geom_histogram(colour = "black") +
  geom_density(col="red") +
  labs(caption = "Figure 8: Book Sales")

p.sales.distribution
```

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

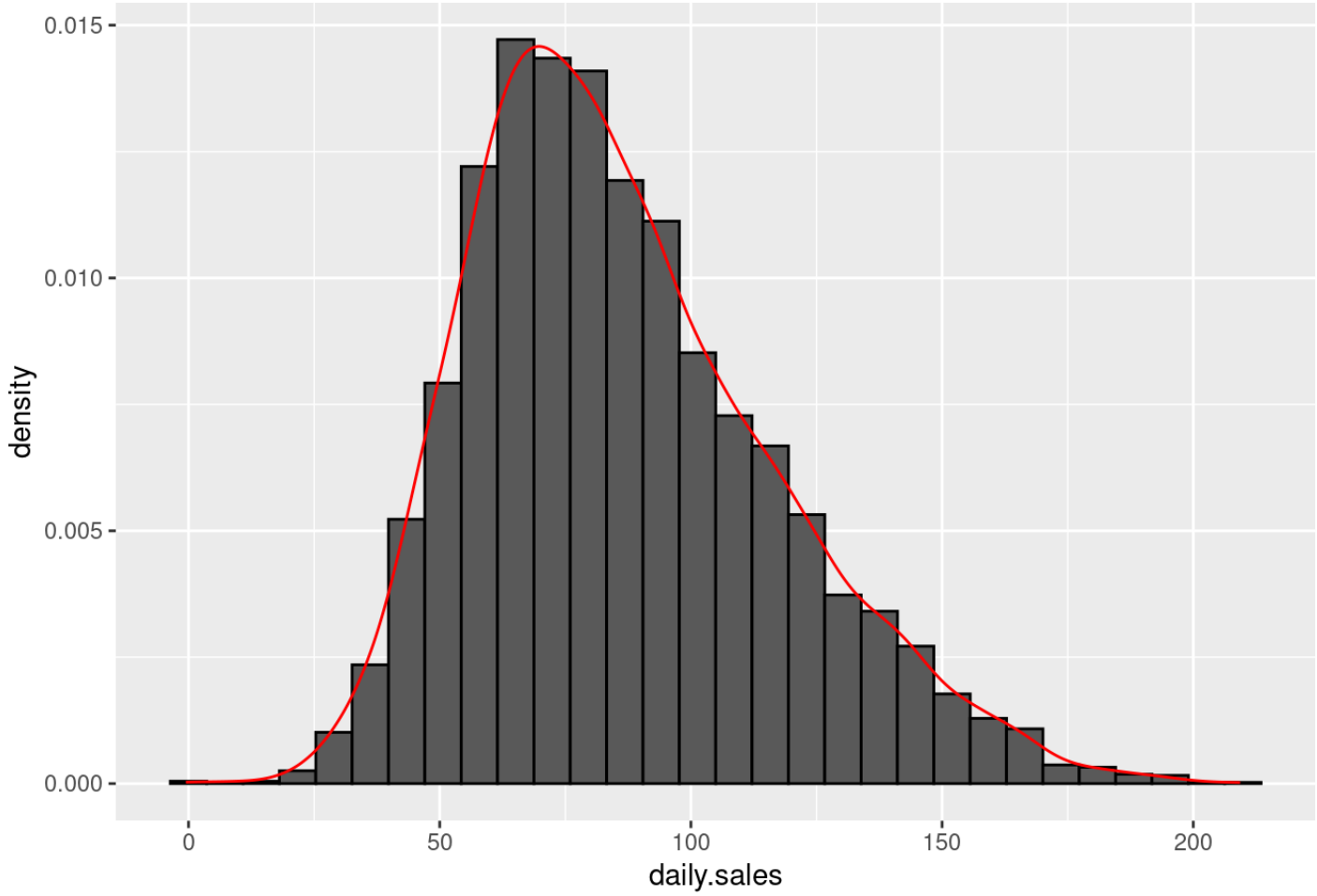


Figure 8: Book Sales

Effect of reviews on the daily sales of the e-book

Books with higher average review scores tend to have higher sales. There is a positive correlation between average review score and daily sales. The relationship between sales and total reviews is less clear. There is no consistent pattern across different review scores.

1. visual representation of relation between the total reviews and average reviews.

```
sales_totalrev_plot <- ggplot(data=sales_data, aes(x=total.reviews, y=daily.sale
s)) +
  geom_jitter(col= 'pink',alpha=0.7)  + geom_smooth()+
  labs(title="Daily sales of the books acoording to their total reviews",x="Total
Reviews",y="Daily Sales",caption = "Figure 9: Daily sales of the books to their to
tal reviews")
sales_totalrev_plot
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Figure 9: Daily sales of the books to their total reviews

```
#the plot shows a positive correlation between total reviews and daily sales.


sales_avgrev_plot <- ggplot(data=sales_data, aes(x=avg.review, y=daily.sales)) +
  geom_jitter(col= 'magenta',alpha=0.7)  + geom_smooth()+
  labs(title="Daily sales of the books acoording to their average reviews",x="Aver
age Reviews",y="Daily Sales",caption = "Figure 10: Daily sales of the books to the
ir average reviews")
sales_avgrev_plot
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Figure 10: Daily sales of the books to their average reviews

```
#the plot shows very weak to no relation between the average reviews and daily sal
es.
```

2. Perform regression analysis to find the best odel to explain the effect on daily sales with respect to tatal reviews and average reviews.

```
model_avgrev <- lm(daily.sales ~ avg.review, sales_data)
summary(model_avgrev)
```

```
##
## Call:
## lm(formula = daily.sales ~ avg.review, data = sales_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -86.721 -22.129  -4.722  18.267 123.403
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  90.8944     2.9543  30.767   <2e-16 ***
## avg.review   -1.0593     0.6859  -1.544    0.123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.22 on 5998 degrees of freedom
## Multiple R-squared:  0.0003974,  Adjusted R-squared:  0.0002308
## F-statistic: 2.385 on 1 and 5998 DF,  p-value: 0.1226
```

```
# The t(5998)= -1.544, p = 0.123. The p-value for the average review coefficient i
s 0.123, which is greater than the commonly used significance level of 0.05.The av
erage review may not be a statistically significant predictor of daily sales in th
is model.

model_totalrev<- lm(daily.sales ~ total.reviews, sales_data)
summary(model_totalrev)
```

```
##
## Call:
## lm(formula = daily.sales ~ total.reviews, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.667  -14.694   -1.125   13.458  147.319
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.381371   1.070717   15.30   <2e-16 ***
## total.reviews  0.527490   0.007761   67.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.72 on 5998 degrees of freedom
## Multiple R-squared:  0.4351, Adjusted R-squared:  0.435
## F-statistic:  4620 on 1 and 5998 DF,  p-value: < 2.2e-16
```

```
# The t(5998)= 67.97,p-value < 0.001. The p-value for the total review coefficient
is highly significant.The total reviews are a statistically significant predictor
of daily sales in this model.

model_rev <- lm(daily.sales ~ avg.review + total.reviews, data = sales_data)
summary(model_rev)
```

```
##
## Call:
## lm(formula = daily.sales ~ avg.review + total.reviews, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.573  -14.529   -0.909   13.669  129.721
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.978718   2.357956  14.410   <2e-16 ***
## avg.review    -4.306466   0.514881  -8.364   <2e-16 ***
## total.reviews  0.533428   0.007749  68.836   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.59 on 5997 degrees of freedom
## Multiple R-squared:  0.4416, Adjusted R-squared:  0.4414
## F-statistic:  2371 on 2 and 5997 DF,  p-value: < 2.2e-16
```

```
# The p-value < 0.001 for both the indicators in this model making them statistica
lly significant and making this model a good fit for our data.

correlation_matrix <- cor(sales_data[c("daily.sales", "avg.review","total.review
s")])
correlation_matrix
```

```
##               daily.sales  avg.review total.reviews
## daily.sales    1.00000000 -0.01993611    0.65961263
## avg.review    -0.01993611  1.00000000    0.09161876
## total.reviews  0.65961263  0.09161876    1.00000000
```

```
# This shows a strong correlation between total reviews and daily sales, and a wea
k correlation between the average reviews and daily sales.
# Also there is a weak correlation between total reviews and average reviews, ther
efore we will not face the problem of multicollinearity in the model as the depend
ent variables are not correlated.

model_intr_rev <- lm(daily.sales ~ avg.review * total.reviews, sales_data )
summary(model_intr_rev)
```

```
##
## Call:
## lm(formula = daily.sales ~ avg.review * total.reviews, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.318  -14.401   -0.873   13.622   94.395
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               76.107846   4.169256  18.255  < 2e-16 ***
## avg.review               -14.673445   0.991327 -14.802  < 2e-16 ***
## total.reviews              0.139838   0.033199   4.212 2.57e-05 ***
## avg.review:total.reviews   0.095620   0.007848  12.184  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.32 on 5996 degrees of freedom
## Multiple R-squared:  0.4551, Adjusted R-squared:  0.4548
## F-statistic:  1669 on 3 and 5996 DF,  p-value: < 2.2e-16
```

```
# The p-value < 0.001 for both the indicators in this model making them statistica
lly significant and making this model a better fit for our data.
```

3.Choosing which model is better for our data.

```
#Choosing which model is better for our data.
model <- lm(daily.sales ~ 1, sales_data)
anova(model, model_intr_rev, model_totalrev, model_rev)
```

```
## Analysis of Variance Table
##
## Model 1: daily.sales ~ 1
## Model 2: daily.sales ~ avg.review * total.reviews
## Model 3: daily.sales ~ total.reviews
## Model 4: daily.sales ~ avg.review + total.reviews
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   5999 5479736
## 2   5996 2985945  3   2493791 1669.240 < 2.2e-16 ***
## 3   5998 3095564 -2   -109620  110.062 < 2.2e-16 ***
## 4   5997 3059870  1     35694   71.677 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*## Anova table shows that all the models are statistically significant (p-value<0.001) except model 4 (daily.sales ~ total.reviews). Model 2 (daily.sales ~ avg.review * total.reviews) and Model 3 (daily.sales ~ avg.review) have highest f- statistic score of (1669.24) and (2501.67) respectively.*

#the ANOVA table provides strong evidence that both avg.review and total.reviews, as well as their interaction, have a significant impact on daily.sales. The model is statistically significant overall.

Effect of different genres and sale price on the daily sales of the e-book

Checking how different genres and sale price of the books have an impact on the daily sales of the books.

```
sales_data$genre <- as.factor(sales_data$genre)
levels(sales_data$genre)
```

```
## [1] "adult_fiction" "non_fiction"   "YA_fiction"
```

```
model_genre <- lm(daily.sales ~ genre, sales_data)
summary(model_genre)
```

```
##
## Call:
## lm(formula = daily.sales ~ genre, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.593  -13.963    0.219   13.563   96.277
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       82.5498     0.4984  165.62   <2e-16 ***
## genrenon_fiction -19.0468     0.7049  -27.02   <2e-16 ***
## genreYA_fiction   30.5136     0.7049   43.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.29 on 5997 degrees of freedom
## Multiple R-squared:  0.4562, Adjusted R-squared:  0.4561
## F-statistic:  2516 on 2 and 5997 DF,  p-value: < 2.2e-16
```

```
cbind(coef(model_genre), confint(model_genre))
```

```
##                                2.5 %    97.5 %
## (Intercept)       82.54984  81.57275  83.52694
## genrenon_fiction -19.04679 -20.42862 -17.66497
## genreYA_fiction   30.51360  29.13177  31.89543
```

```
# The intercept shows the daily price for adult fiction [82.55] 95% CI [81.57275 -
- 83.52694], For every one unit increase in non fiction books, daily sales is drop
ped by 19 units 95% CI [-20.42 -- -17.66] and for every one unit increase in young
adult fiction boks, daily sales is increased by 30.51 units 95% CI [29.13 -- 31.8
9]

model_sale <- lm(daily.sales ~ sale.price, sales_data)
summary(model_sale)
```

```
##
## Call:
## lm(formula = daily.sales ~ sale.price, data = sales_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.224 -17.382  -1.922  15.227 113.674
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 127.31696    0.95722  133.01   <2e-16 ***
## sale.price   -3.97625    0.08704  -45.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.03 on 5998 degrees of freedom
## Multiple R-squared:  0.2581, Adjusted R-squared:  0.258
## F-statistic:  2087 on 1 and 5998 DF,  p-value: < 2.2e-16
```

```
cbind(coef(model_sale), confint(model_sale))
```

```
##                              2.5 %    97.5 %
## (Intercept) 127.316964 125.440463 129.19346
## sale.price   -3.976249  -4.146878  -3.80562
```

```
#For one unit increase in sale price, daily sales is dropped by 4 units 95% CI [-4.14 -- -3.80]
```

```
sales_genre_plot <- ggplot(data=sales_data, aes(x=genre, y=daily.sales)) +
  geom_violin(trim=FALSE,fill= 'pink',alpha=0.5) +
  stat_summary(fun.data = mean_sdl,geom="pointrange",col="blue",fun.args=list(mult=1)) +
  labs(title="Daily sales of the books for different genre",subtitle =" Error Bars are Standard Deviation",x="Genre",y="Daily Sales",caption = "Figure 11. Daily sales of the books for different genre")
sales_genre_plot
```

Figure 11. Daily sales of the books for different genre

```
# The plot show that the mean of daily sales is highes for young adult fiction.

saleprice_dailysales_plot <- ggplot(data=sales_data, aes(x=sale.price, y=daily.sal
es)) +
  geom_jitter(col= 'pink',alpha=0.7)  + geom_smooth()+
  labs(title="Daily sales of the books acoording to their sale price",x="Sale Pric
e",y="Daily Sales",caption = "Figure 12: Daily sales of the books with their Sale
Price")
saleprice_dailysales_plot
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Figure 12: Daily sales of the books with their Sale Price

```
#The plot does not show a very clear relation between sale price and daily sale, b
ut we can infer that as the sale price increase daily sales tend to drop.

saleprice_dailysales_plot_genre <- ggplot(data=sales_data, aes(x=sale.price, y=dai
ly.sales)) + facet_wrap(~genre)+
  geom_jitter(col= 'violet',alpha=0.7)  + geom_smooth()+
  labs(title="Daily sales of the books according to there genres with their Sale P
rice",x="Sale Price",y="Daily Sales",caption = "Figure 13: Daily sales of the book
s according to there genres with their Sale Price")
saleprice_dailysales_plot_genre
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Figure 13: Daily sales of the books according to there genres with their Sale Price

```
#The plot shows the effect of sales price of different genres on the daily sales.
# we can see that adult fiction has a constant relation till the sales price of 10
and then there is a negative correlation between daily sales and sale price as it
increases, for non fiction sales price have no to very weak impact on the daily sa
les, and for young adult fiction sale price has a strong negative correlation with
daily sales.
```

```
m.sales.by.genre <- lm(daily.sales~ sale.price + genre, data=sales_data)
summary(m.sales.by.genre)
```

```
##
## Call:
## lm(formula = daily.sales ~ sale.price + genre, data = sales_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -114.304  -13.736    0.289   13.548   96.295
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       93.9570     1.2437  75.546  < 2e-16 ***
## sale.price        -1.4311     0.1432  -9.996  < 2e-16 ***
## genrenon_fiction  -9.0252     1.2223  -7.384 1.75e-13 ***
## genreYA_fiction   30.4787     0.6992  43.594  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.11 on 5996 degrees of freedom
## Multiple R-squared:  0.4651, Adjusted R-squared:  0.4649
## F-statistic:  1738 on 3 and 5996 DF,  p-value: < 2.2e-16
```

```
# The intercept shows the daily price for adult fiction [93.95],for every one unit increase in the sale price, daily sales is dropped by 1.43, For every one unit increase in non fiction books, daily sales is dropped by 9.02 units and for every one unit increase in young adult fiction books, daily sales is increased by 30.47 units.

anova( model,model_sale, m.sales.by.genre, model_genre)
```

```
## Analysis of Variance Table
##
## Model 1: daily.sales ~ 1
## Model 2: daily.sales ~ sale.price
## Model 3: daily.sales ~ sale.price + genre
## Model 4: daily.sales ~ genre
##   Res.Df     RSS Df Sum of Sq        F    Pr(>F)
## 1   5999 5479736
## 2   5998 4065261  1   1414475 2893.778 < 2.2e-16 ***
## 3   5996 2930838  2   1134423 1160.419 < 2.2e-16 ***
## 4   5997 2979674 -1    -48836   99.911 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Anova table shows that all the models are statistically significant (p-value<0.0
01) but model 2 (daily.sales ~ sale.price) and model 3 (daily.sales ~ sale.price +
genre) both have a very high f-statistic of (2893.78) and (1160.42) respectively.

#plotting our findings.
( m.sales.by.genre.emm <- emmeans(m.sales.by.genre, ~sale.price + genre)  )
```

```
##  sale.price genre         emmean    SE   df lower.CL upper.CL
##        10.3 adult_fiction   79.2 0.596 5996     78.1     80.4
##        10.3 non_fiction     70.2 0.832 5996     68.6     71.8
##        10.3 YA_fiction     109.7 0.598 5996    108.5    110.9
##
## Confidence level used: 0.95
```

```
ggplot(summary(m.sales.by.genre.emm), aes(x=genre, y=emmean, ymin=lower.CL, ymax=u
pper.CL, col=genre)) + geom_point() + geom_linerange(alpha=0.5) + labs(title=" Est
imated mean of daily sales of books for genre",x="genre", y="Daily Sales", col="Ge
nre", subtitle="Error bars are 95% CIs", caption = "Figure 14: Estimated mean of d
aily sales of books for genre")
```

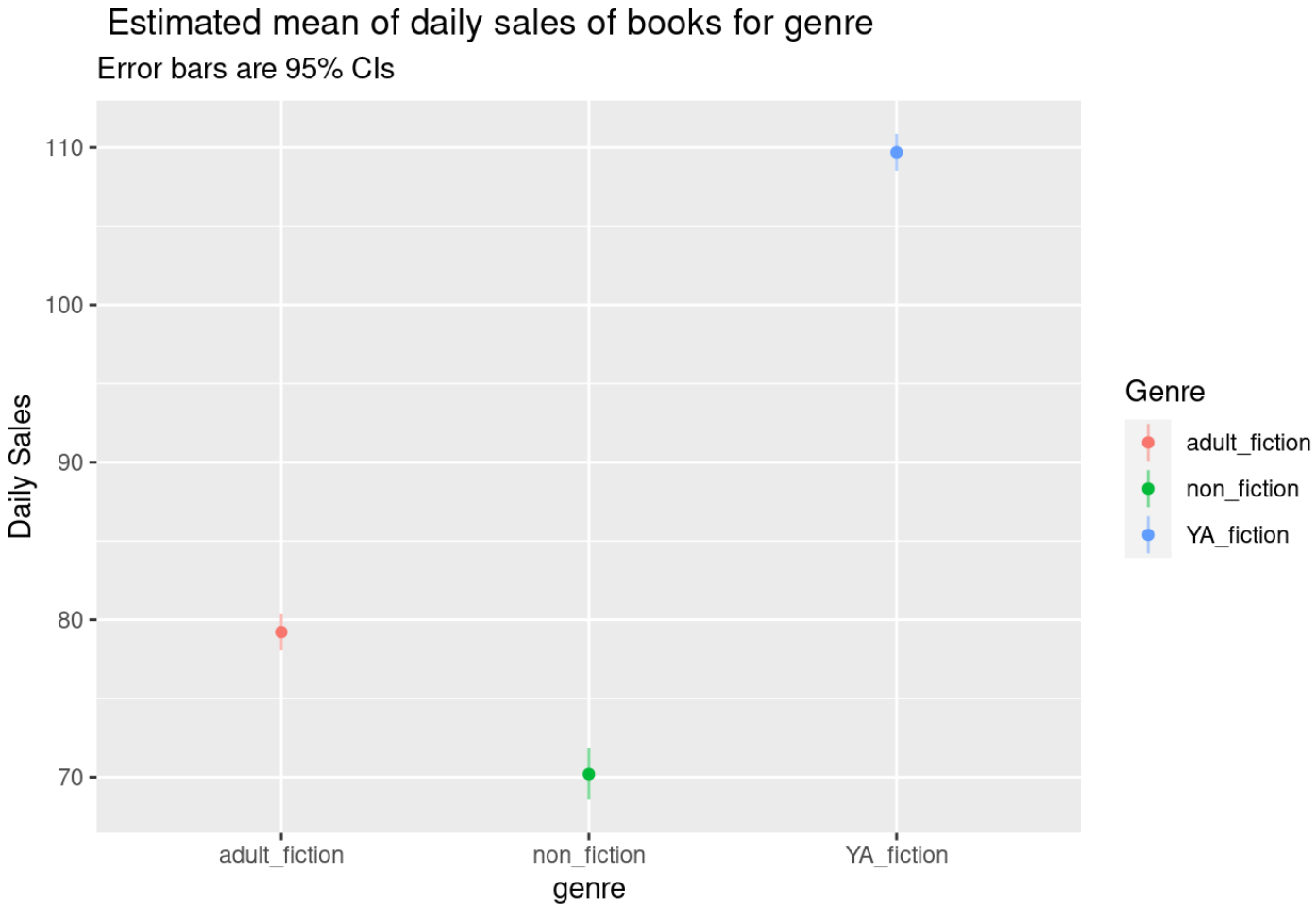


Figure 14: Estimated mean of daily sales of books for genre

Report on Daily Sales of e-book

This report presents the results of the analyses requested by the company. The data provides information on e-books sales over a period of time containing 6000 observations.The analysis focuses on exploring the impact of various factors on e-book sales over several months. The key variables of interest are the average review scores, total reviews, genre, and sale prices.

Upon loading the data set, it was subjected to a comprehensive data quality check. The data set contains information on the publisher, publisher type, genre, average review, daily sales, total reviews, and sale price for each book. No duplicates were found, and missing values were minimal.

A jitter plot was utilized to visually assess the distribution of daily sales. The distribution appeared approximately normal, and a histogram with a density plot provided additional insights into the overall sales pattern.

Books with higher Total review scores exhibit a positive correlation with daily sales. Regression analysis indicated that total reviews significantly predict daily sales, whereas the relationship with average reviews was less conclusive.

Figure 9: Daily sales of the books to their total reviews

It can be seen that there a positive correlation between the books with higher total reviews with dail sales.

Figure 10: Daily sales of the books to their average reviews

This shows that the correlation between average reviews and daily sales is very weak.

A regression model incorporating both average and total reviews, along with their interaction, emerged as the most suitable with F-statistic: 2371 on 2 and 5997 Degrees of Freedom, p-value: < 2.2e-16.

Regression models were compared using ANOVA table, highlighting the superiority of the model incorporating both average and total reviews along with their interaction.The selected model demonstrated statistical significance and accounted for a significant proportion of the variance in daily sales. F- Statistic 1669.240, p-value < 2.2e-16.

I then looked at how differences in book genres demonstrated varying effects on daily sales.

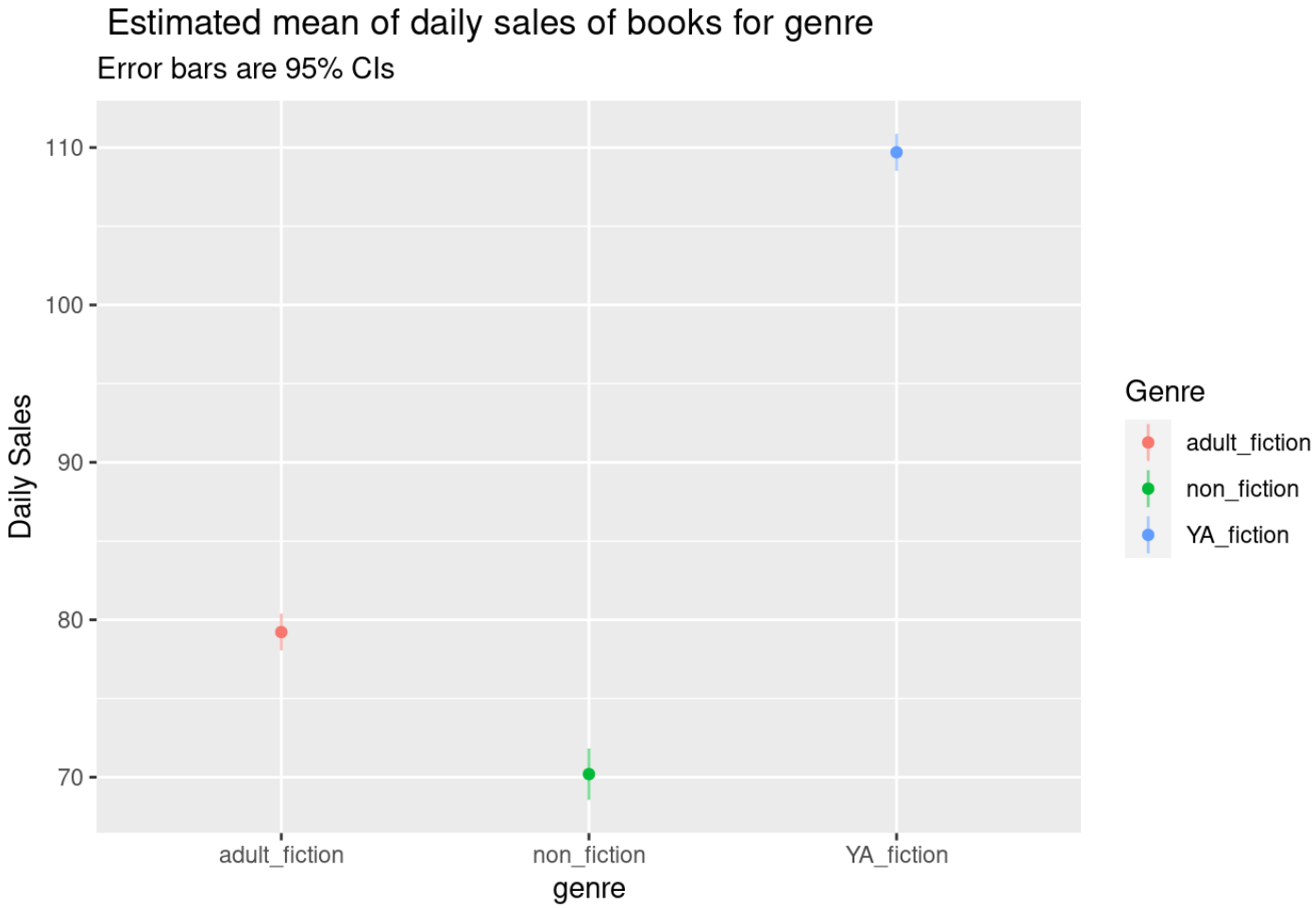


Figure 14: Estimated mean of daily sales of books for genre

Where I found that Young adult fiction had the most substantial impact, leading to increased daily sales. Then the adult fiction and non fictional books had least impact on the daily sales of e-book.

Also I looked at the effect of sale price of different genre on the daily sales of e-book

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Figure 13: Daily sales of the books according to there genres with their Sale Price

This shows that adult fiction has a constant relation till the sales price of 10 and then there is a negative correlation between daily sales and sale price as it increases, for non fiction sales price have no to very weak impact on the daily sales, and for young adult fiction sale price has a strong negative correlation with daily sales.

Sale prices displayed a negative relationship with daily sales. A higher sale price was associated with decreased daily sales. For one unit increase in sale price, daily sales is dropped by 4 units 95% CI [-4.14 – -3.80].

The analysis reveals that both reviews and genres significantly influence daily sales of e-books. Young adult fiction and lower sale prices contribute positively to sales. Recommendations for publishers could involve strategies to encourage positive reviews, particularly for genres with higher sales potential. Continuous monitoring and adaptation of pricing strategies may optimize daily sales.

Limitations and Future Work: The analysis assumes linearity and normality in the relationships, and the results are based on correlational findings. Future work could involve exploring additional variables, investigating temporal trends, and employing advanced modeling techniques for a more nuanced understanding.

References

1.R for Data Science (https://r4ds.had.co.nz (https://r4ds.had.co.nz))

2.G. Cumming. ‘The New Statistics: Why and How’,https://journals.sagepub.com/doi/10.1177/0956797613504966 (https://journals.sagepub.com/doi/10.1177/0956797613504966)

3.https://data.london.gov.uk/dataset/number-bicycle-hires

4.https://data.london.gov.uk/dataset/covid-19-restrictions-timeseries