STAT 4003 Final Project

Coles, Mai, Mehrabi, Tabler

December 12th 2018

INTRODUCTION

We chose to source two data sets, Los Angeles Parking Citations from Los Angles Open Data and Marvel/DC Comic Data from Tidy Tuesday, do to their large observation count and their ability as a data set to be manipulated. The LA Parking Citations summate parking tickets given within Los Angeles, California City Limits from December 2015 through June 2016, totaling over four billion tickets. There are nineteen recorded variables for all of the tickets given, which are listed in the data summary. The Marvel/DC Comic Data includes descriptive characteristics of over twenty three thousand comic books characters over the two comic universes from 1935 through 2013. Concerning these two data sets, we narrowed our scope to two specific questions for each respective set.

Does the make and color of a vehicle result in a higher proportion of fines administered by the city of LA? How does the time of day effect the amount of tickets administered? Both of these questions led to the following four hypotheses. Our first null and alternate hypotheses concerning the color of vehicles in Los Angeles are as follows, respectively. Null: There is no significance between the color of vehicle and the proportion of fines that color vehicle is administered. Alternate: There is statistical significance between the color of a vehicle and the proportion of fines that color vehicle receives. The hypotheses that fit our question for the time of day at which tickets are administered in Los Angeles proceed. Null: There is no significance between the time of day, month or year and the amount of fines received. Alternate: There is statistical significance between the time of day, the month, or the year and the amount of fines received in Los Angeles.

link to parking ticket data: here

Note: Since the data set consists of a few billion data points we sampled 100000 data points so we would have a manageable data size and still having a large data set that represents our data.

```
#n=100000
#dat<-read.csv(here('Final Project','parking-citations.csv'))
#smalldat<- dat[sample(1:nrow(dat), n, replace=FALSE),]
#write.csv(smalldat,file=here('Final Project','parking-sub-dat.csv'))
```

Concerning the Marvel/DC Comic Data, we wanted to know how social construct plays into the type of characteristics given to characters in the two major comic universes. Our first question intendeds to reveal how each universe incorporates gender equity over time. How does the proportion of female appearances, by universe, compare to that of males as the years progress? Secondly, we ask: Is there evidence to suggest that darker physical features (brown eyes, dark hair) correspond to a villain or 'bad character,' as opposed to a good one? These two questions led to the follow null and alternate hypotheses. Null: There is no significant difference in the levels of gender equity between the two comic universes, Marvel and DC. Alternate: There is statistical significant difference between Marvel and DC, when looking into their gender equity over time. Secondly, Null: There is not a suitable statistical significance between the hair or eye color of a character and their alignment (good or bad character). Alternate: When it comes to hair and eye color and the alignment of a character there is a lack of independence between the two.

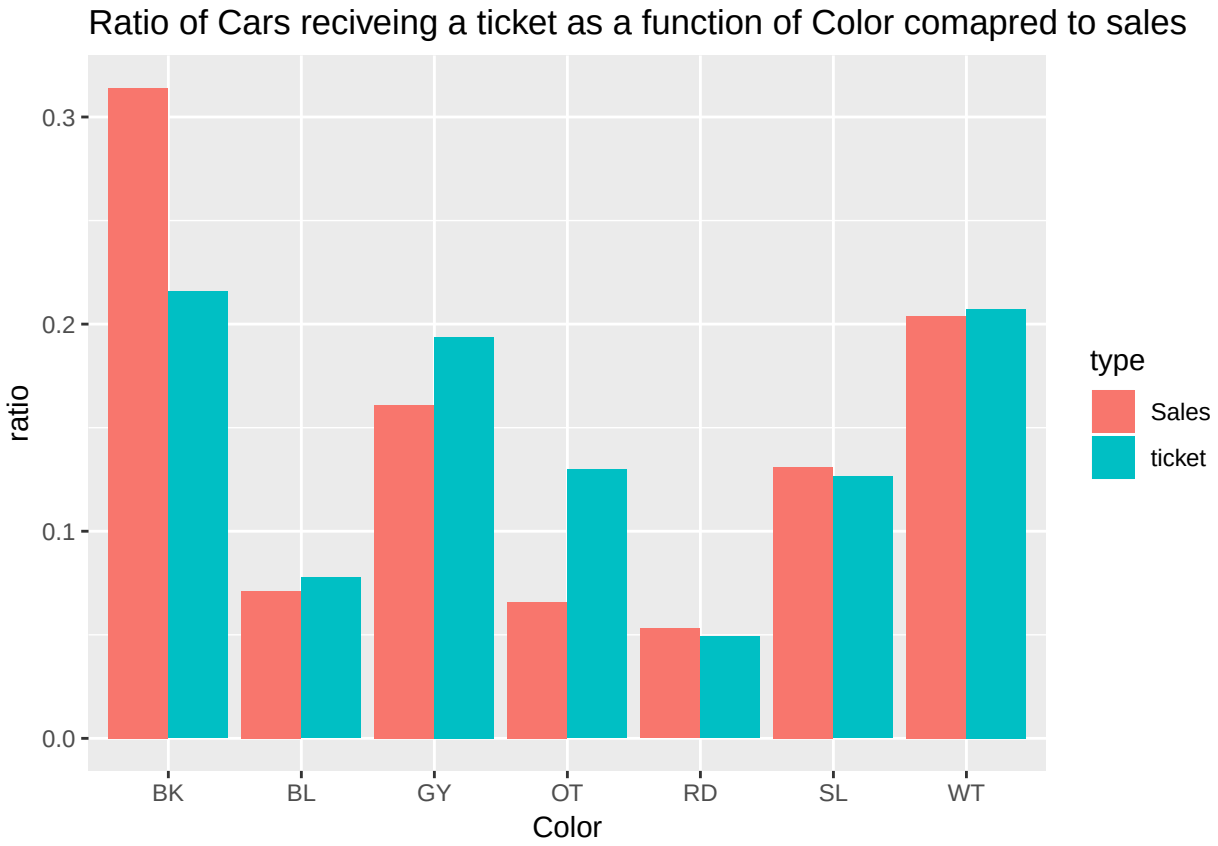
link to comicbook characters data: here

EXPLORATORY DATA ANALYSIS

After tackling the problem of visualizing how much and when tickets were administered in Los Angeles, we then had to isolate the tickets by the color of the vehicle they were written to. To do this, we created a histogram of our data subset, with the number of vehicles receiving a ticket as a function of color. After noting that six of the most common colors dominated the rest of the colors we compiled the remaining colors into a category called 'Other.' In making this graph, we were able to then compare 2015's color percentages to that of the ticketed vehicles in Los Angeles. After comparing these, we ran a chi squared test to determine if the tickets given were independent of the color as they pertain to national statistics.

link to Popular car colors of 2015: here

```
##
##  Chi-squared test for given probabilities
##
## data:  TicketCarCol$number
## X-squared = 64231, df = 6, p-value < 2.2e-16
```



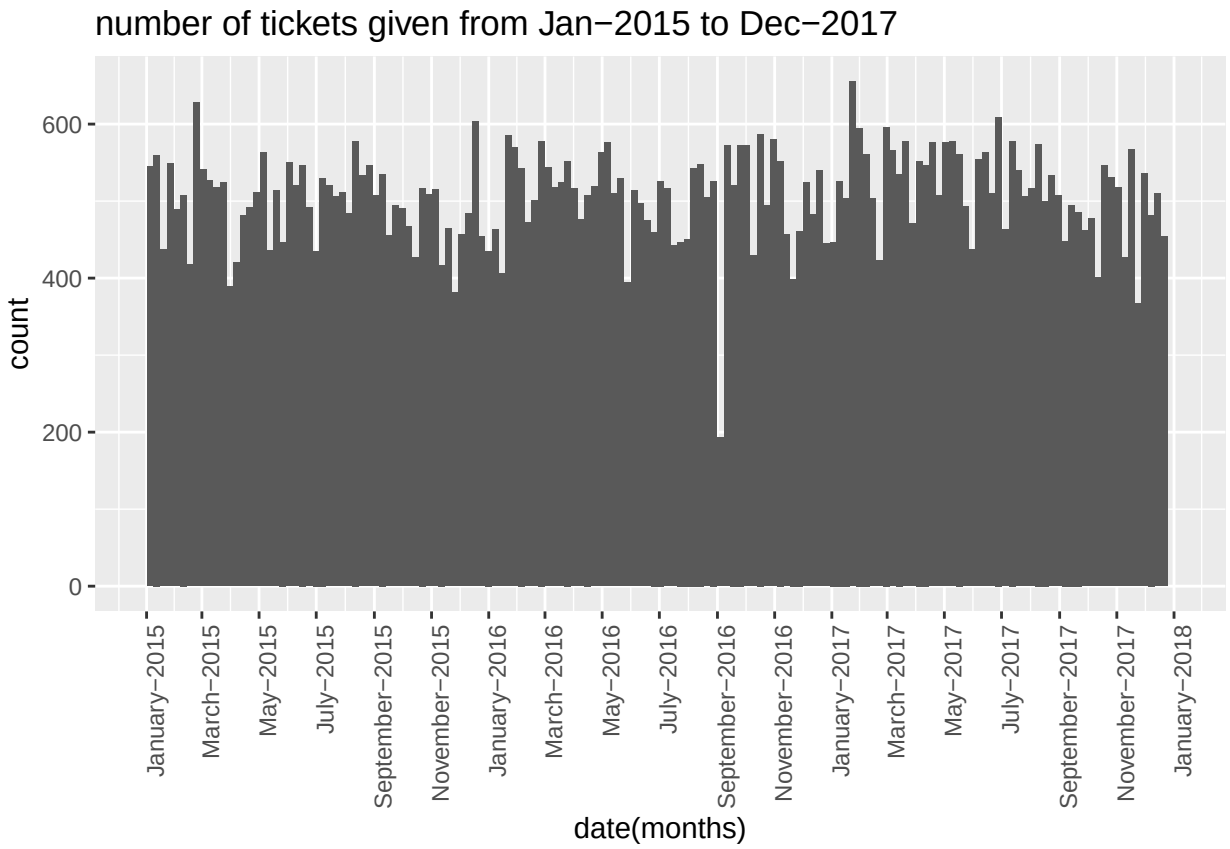
Note: BK, BL, GY, OT, RD, SL, WT stand for Black, Blue, Gray , Other, Red , Silver and White respectively.

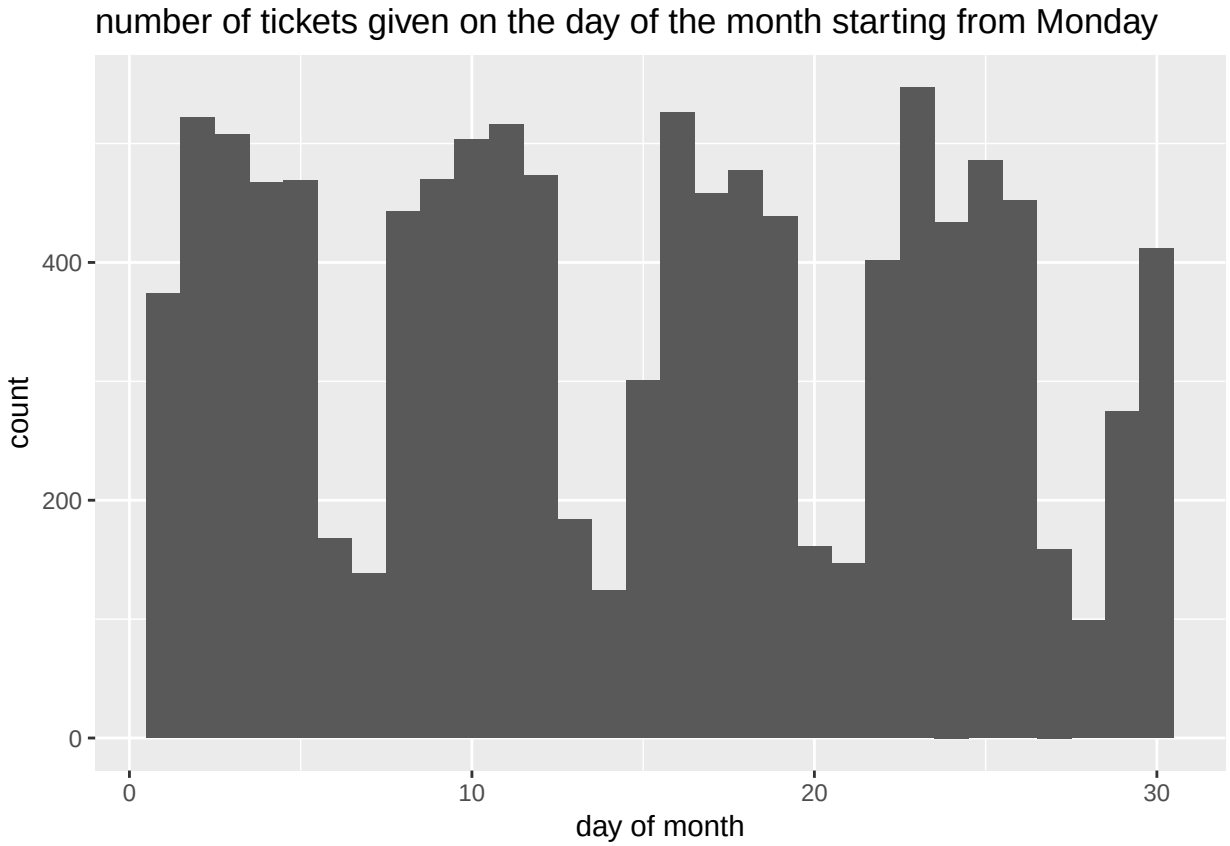
When first delving into our data set, Los Angeles Parking Citations, we encountered a large number of observations with nineteen recorded variables. These different variables ranged from the color of the car ticketed to the location at which the ticket was given. When faced with such a wide range of variables, it was important to narrow our scope. We decided to look into the rates at which tickets were given over time as well as if different colors of vehicles were getting ticketed disproportionately. To accomplish this, we first created a subset of the data to get a look at a more manageable number of observations. After creating this subset, we wanted to visualize the data in two different ways, tickets given over time and tickets given to a specific color of vehicle. To visualize administered tickets over time, we built graphs of the amount of tickets

as they were given throughout hours of the day, days of the month, and months of the year. What we were able to create was trend map showing when were the most and least frequented times for tickets, how many tickets were given, and were even then able to infer the causes for highs and lows of parking tickets in Los Angeles.

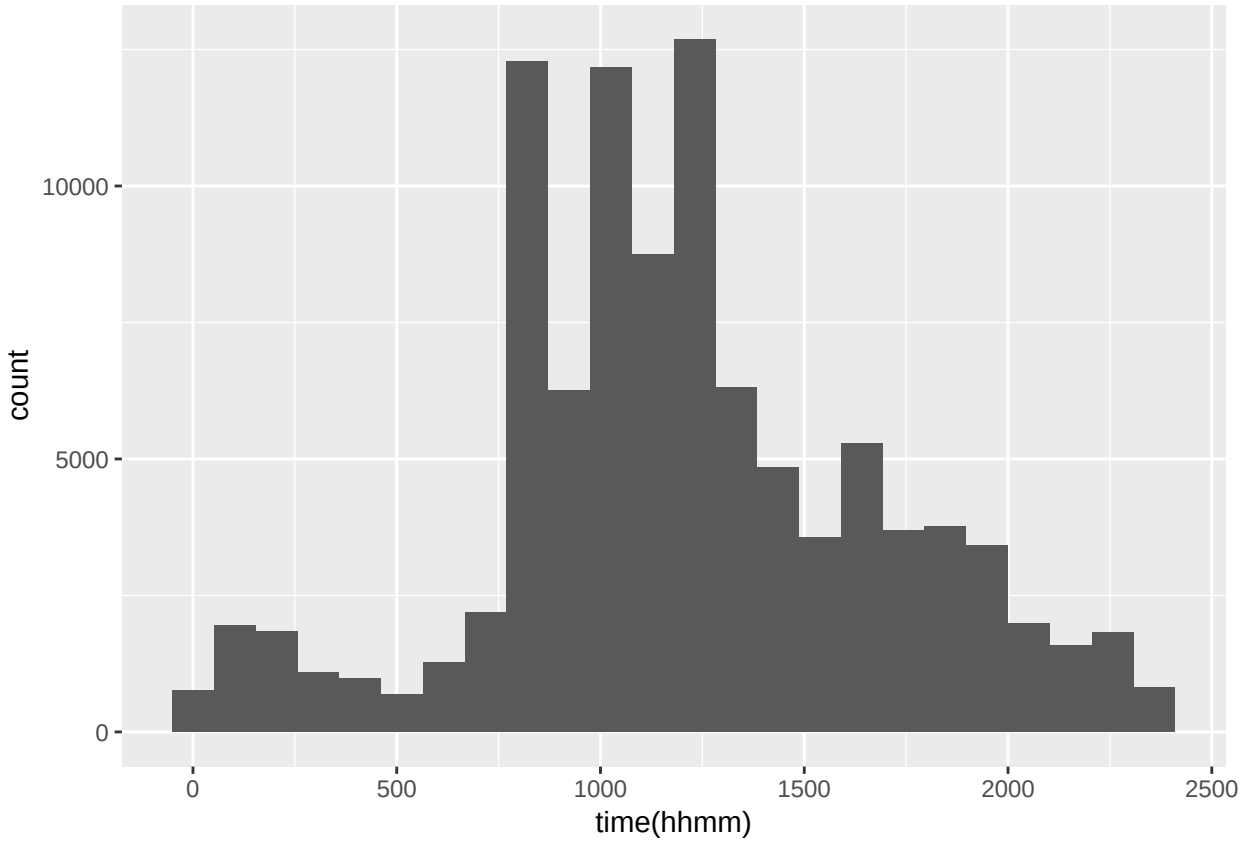
```
## Warning: Removed 21555 rows containing non-finite values (stat_bin).
```

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

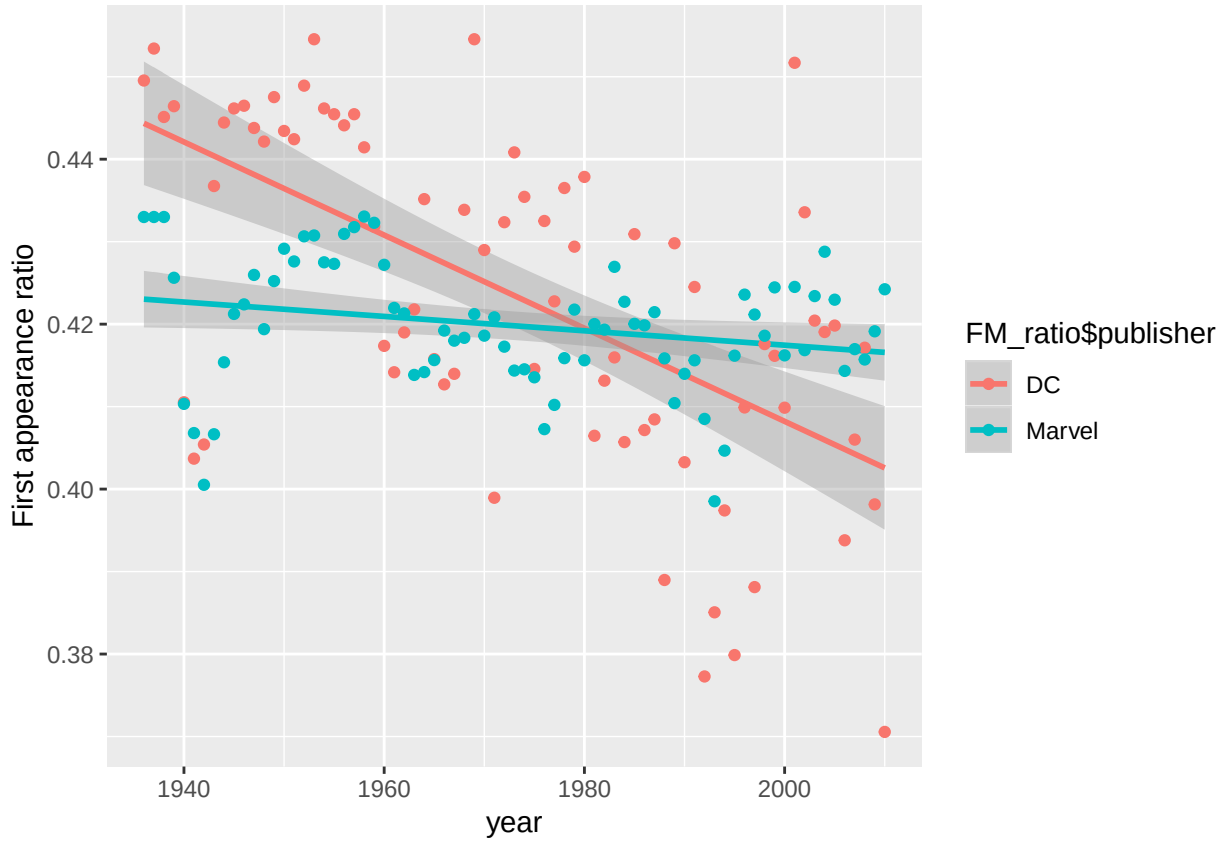




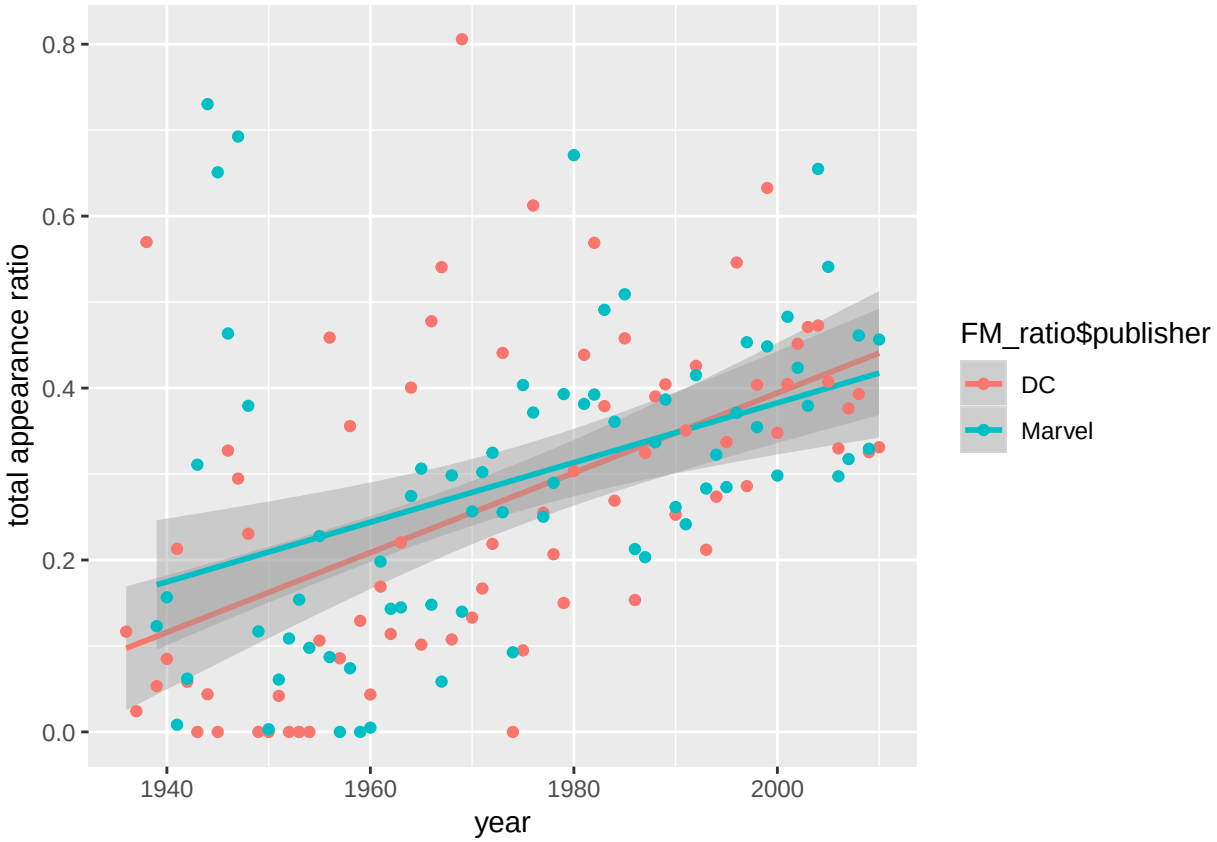

```
## Warning: Removed 25 rows containing non-finite values (stat_bin).
```



Our second data set, which included characters from the Marvel/DC universes, posed more topical questions. Our first question was interested in how the proportion of female characters compares to male characters as the comic universes have progressed and more characters have been created. We have data that recorded the first appearance of each character in both the Marvel and DC universes since 1935. Using this, we initially created a histogram of the number of newly introduced characters in both the DC and Marvel universes, with each bar being divided into male and female characters. What we could initially tell from these graphs is that there are much more characters being created now than there were. Additionally, more male characters are created most years than female characters. To understand the gender equity of the comic universes more, we ran a linear regression model for both universes of the total female appearances over the years, as well as, the ration of female appearances over the years. What we then formatted was two graphs with a first appearance ratio over the years and a total appearance ratio over the year. These show that both Marvel and DC comics have an insignificant difference in the proportion of female characters that are appearing in their comics; however, Marvel creates and introduces and significantly different amount of female characters than DC, proportionately.




```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
## Warning: Removed 3 rows containing missing values (geom_point).
```



Secondly, we wanted to know whether or not physical characteristics where independent of the character's alignment, or not. To understand this, we ran a chi squared test. In using the chi squared test of independence, we tested hair color and eye color against the character's alignment. When testing both the hair and eye color, we found that there were only eight and nine significant sub-variables, respectively. In the set for hair, there were observations for both 'Bald' and 'No Hair,' so they were then combined. After running both of the chi squared tests at a 99% confidence interval, it was supported that hair nor eyes were independent of a character's alignment in either of the Marvel or DC universes.

```
## Warning in chisq.test(Chi_hair$hair, Chi_hair$align): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Chi_hair$hair and Chi_hair$align
## X-squared = 487.21, df = 12, p-value < 2.2e-16
```

```
## Warning in chisq.test(Chi_eye$eye, Chi_eye$align): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Chi_eye$eye and Chi_eye$align
## X-squared = 386.1, df = 16, p-value < 2.2e-16
```

DATA ANALYSIS

To test our null hypothesis stating that there was no significance between the color of vehicle and the proportion of fines that colored vehicles were administered, we developed a graph of the number of vehicles ticketed and the proportion of vehicles purchased by the color of the vehicle. After seeing that black vehicles were ticketed at a much lower number than what was expected by national car sales, a Chi Squared test was ran at a confidence level of 99%, and reported back a p-value of $2 ^{-16} $. From this p-value, it can be supported that numbers of tickets and the color of the vehicle ticketed are not independent. What does this mean? In our graph it was shown that black vehicles were ticketed at a lower rate, and 'other' colored vehicles were ticketed at a higher rate than expected. This could be due to a number of factors. With the installation of Uber and UberX in bigger cities, most of which are black, parking tickets should not usually be an issue. 'Other' colored cars could also be flashy and draw attention to an illegally parked car. Also, black cars may not be as proportionately owned in Los Angeles due to the heat, meaning that the number of parking tickets would be lower. This last explanation is one that is most feasible, leading us to believe there is discrepancy in car ownership by color, rather than ticketing. We did, in our analysis, make the assumption that car colors in Los Angeles would fit closely to the national average. This is a weakness in our data, and could be an overcome obstacle if we could determine the car color percentages in Los Angeles, specifically. Our next null hypothesis involving the Los Angeles Parking Citations data states that there is no significance between the time of day, month, or year and the amount of fines that are received. After analyzing and restructuring the data by day, month and year, we were able to create a time map of when and at what rate tickets were being given in the Los Angeles city limits. We observed that ticketing was low from 8pm to 5:30am. When looking at our next graph we see that ticketing dies down on Saturday and Sunday by close to half. Finally, we see a steady trend hovering around 500 tickets a week, for most months of the year. What we have found is that the parking tickets are given on a strict schedule, sticking to a definite trend with only a few days out of the year deviating from the norm. One notable day with less tickets being September 11th, a day of national mourning.

We then set out to test our third and fourth hypotheses, coming from our Comic Data. Our first hypothesis we tested with a linear regression. Our null hypothesis stated that there is no significant difference in the levels of gender equity between the two comic universes in our data set. There was access to over twenty-three thousand characters, which we used to test this gender equity. After discovering how the introduction of female characters into both of the universes had increased over the years, it was important for us to run a linear model comparing the ratios of the first appearance of the characters and total appearances of the characters. The test showed, that within our error margin, there was no significant difference in the ratio of total female appearances between the two comic universes. However, as we were surprised to find, there was a difference in the ratio of first female appearances between the two. This was supported by our p values of 0.003 and 1.12×10^{-8} for Marvel and DC respectively when testing the ratio of first female appearances, and 3.17×10^{-4} and 5.35×10^{-7} for Marvel and DC respectively, when testing the total appearance ratio. What this means is that DC develops their female characters more, by having more total female appearances, per the amount of their first female appearances. Our data was strong in that we were able to draw from a large around of observations; however, there were some observations that could not be used due to lack of a variable, such as 'first appearance date.' A more comprehensive list of comic characters, even those outside of the Marvel/DC universes, would be interesting to tie into this. Lastly, we asked whether or not the physical features of comic characters are independent of their alignment, or whether they are a 'good' or 'bad' character. Our null hypothesis reads as follows: There is not a statistical significant difference between the hair or eye color of a character and their alignment. It could be said that the physical characteristics of our comic heroes and villains should not have an effect on how they are portrayed in the books. To determine this we formulated a Chi Squared Test that ran both hair and eye color against the alignment of the twenty-three thousand characters. After removing observations whose variable totals were insignificant to that of the total population, we were able to get a p-value for both of our Chi Squared tests that showed significant differences in our variables at a 99% confidence interval. Both tests had a p-value of nearly zero, and were calculated at 2.2*10^-16 for both eye color and hair color against character alignment. Based on this we know that a character's physical traits play some role in their alignment. This, when looking at the data, is mainly due to the fact that bald characters were most typically villains, and those with blonde hair often good characters. Characters with blue eyes were found to be overwhelmingly good, and those with red or

yellow eyes were overwhelmingly bad. Our data was strong in this field, being that eye and hair color was recorded for most all characters, but there was the problem of some of the variable options not fitting in the test statistic. Options like, 'photocellular eyes.' Altogether, it still stands that it is possible to predict the alignment of the character, given their physical characteristics.

CONCLUSION

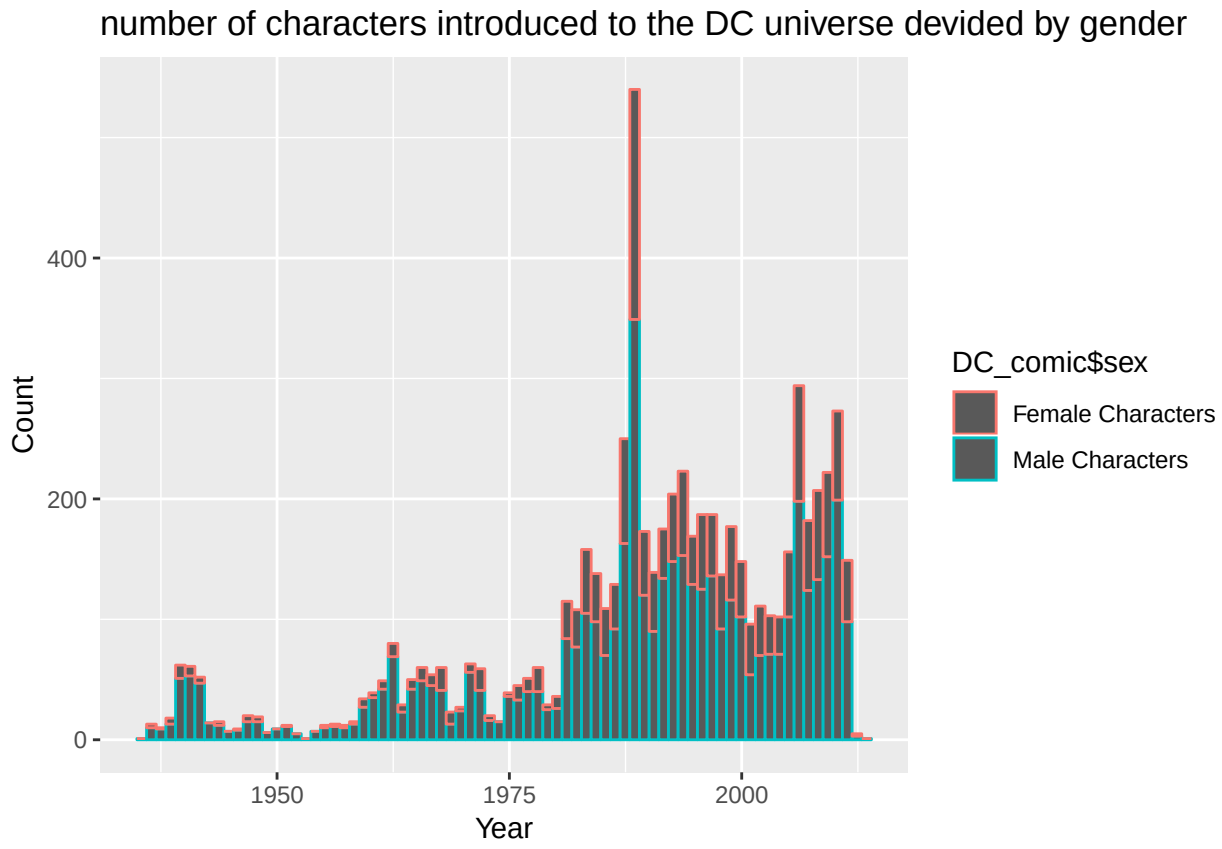
After analyzing the two data sets, Los Angeles Parking Citations and Marvel/DC Comic Data, asking our four questions, and formulating our hypotheses, this is what has been compiled. It's shown that in Los Angeles black cars are being ticketed at a lower rate than expected, whereas, collectively, vehicles that have a color other than the six most popular are being ticketed at a higher rate than expected. This is supported by our Chi Squared test. Additionally, it is shown that the tickets administered to these Los Angeles cars fall within mid-day, during weekdays, consistently over the course of a year. The total ticket values in Los Angeles over the course of a day, month, or year rarely deviate from what is normal. As for our Marvel/DC Comic Data, we found in each of our questions a lack of independence in some way. When looking at the gender equity of comic characters, we found that there was a significant difference in the ratio of female first appearances over the years, between DC and Marvel characters. What we concluded from our investigation into this data is that DC invests more comic appearances on already developed female characters, while female Marvel characters, on average, do not make as many appearances. This, in addition to our findings on the lack of independence between hair and eye color and character alignment, show that Marvel and DC comics have a different approach to how the describe their describe and depict their characters. Based on our Chi Squared test we can support the fact physical traits are strongly tied to whether a character is good or bad.

Appendix

Histogram of ratio of male and female characters over the years for Marvel and DC

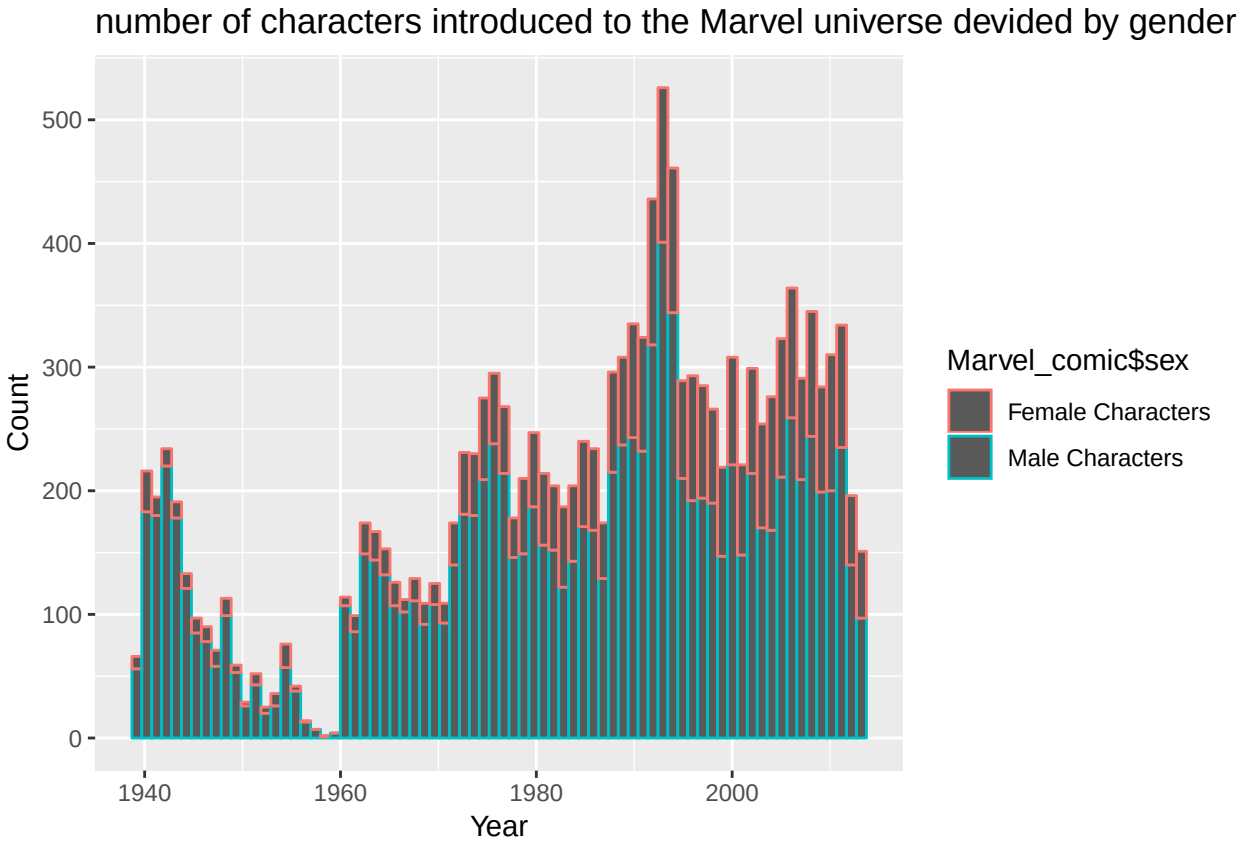
```
datcomic<-read.csv(here('Final Project','tidytuesday-comic.csv'))
datcomic<-datcomic[datcomic$sex!= 'Genderless Characters'& datcomic$sex!= 'Agender Characters'& datcomi

DC_comic<-datcomic[datcomic$publisher=='DC',]
Marvel_comic<-datcomic[datcomic$publisher=='Marvel',]

ggplot(data=DC_comic ,aes(x=DC_comic$year,color=DC_comic$sex))+
geom_histogram(bins=76)+
xlab('Year')+
ylab('Count')+
ggtitle('number of characters introduced to the DC universe devided by gender')
```

```
## Warning: Removed 1047 rows containing non-finite values (stat_bin).
```




```
ggplot(data=Marvel_comic ,aes(x=Marvel_comic$year,color=Marvel_comic$sex))+
geom_histogram(bins=74)+
xlab('Year')+
ylab('Count')+
ggtitle('number of characters introduced to the Marvel universe devided by gender')
```

```
## Warning: Removed 1726 rows containing non-finite values (stat_bin).
```



Linear models for female ratio of first appearance anrd total number of appearance in DC and marvel:

```
##linear model  of  the female total appearances over the years in The DC universe
DC_app_fit<-lm(DC_FM_ratio_df$app_ratio~DC_FM_ratio_df$year)
summary(DC_app_fit)
```

```
##
## Call:
## lm(formula = DC_FM_ratio_df$app_ratio ~ DC_FM_ratio_df$year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27363 -0.10928 -0.04427  0.06614  0.55536
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -8.8869702  1.6650636  -5.337 1.02e-06 ***
## DC_FM_ratio_df$year  0.0046406  0.0008439   5.499 5.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1582 on 73 degrees of freedom
## Multiple R-squared:  0.2929, Adjusted R-squared:  0.2832
## F-statistic: 30.24 on 1 and 73 DF,  p-value: 5.346e-07
```

```r
##linear model  of  the female total appearances over the years in The Marvel universe
Marvel_app_fit<-lm(Marvel_FM_ratio_df$app_ratio~Marvel_FM_ratio_df$year)
summary(Marvel_app_fit)
```

```
##
## Call:
## lm(formula = Marvel_FM_ratio_df$app_ratio ~ Marvel_FM_ratio_df$year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24049 -0.10817 -0.01964  0.06583  0.54188
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -6.5566421  1.8086050  -3.625 0.000544 ***
## Marvel_FM_ratio_df$year  0.0034697  0.0009159   3.788 0.000318 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1615 on 70 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.1701, Adjusted R-squared:  0.1583
## F-statistic: 14.35 on 1 and 70 DF,  p-value: 0.0003178
```

```r
##linear model  of  the female  appearance ratio by year in The DC universe
Marvel_ratio_fit<-lm(Marvel_FM_ratio_df$ratio~Marvel_FM_ratio_df$year)
summary(Marvel_ratio_fit)
```

```
##
## Call:
## lm(formula = Marvel_FM_ratio_df$ratio ~ Marvel_FM_ratio_df$year)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0219862 -0.0035343  0.0005367  0.0059456  0.0119365
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.918e-01  7.920e-02   7.472 1.37e-10 ***
## Marvel_FM_ratio_df$year -8.719e-05  4.014e-05  -2.172   0.0331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007526 on 73 degrees of freedom
## Multiple R-squared:  0.0607, Adjusted R-squared:  0.04783
## F-statistic: 4.717 on 1 and 73 DF,  p-value: 0.03311
```

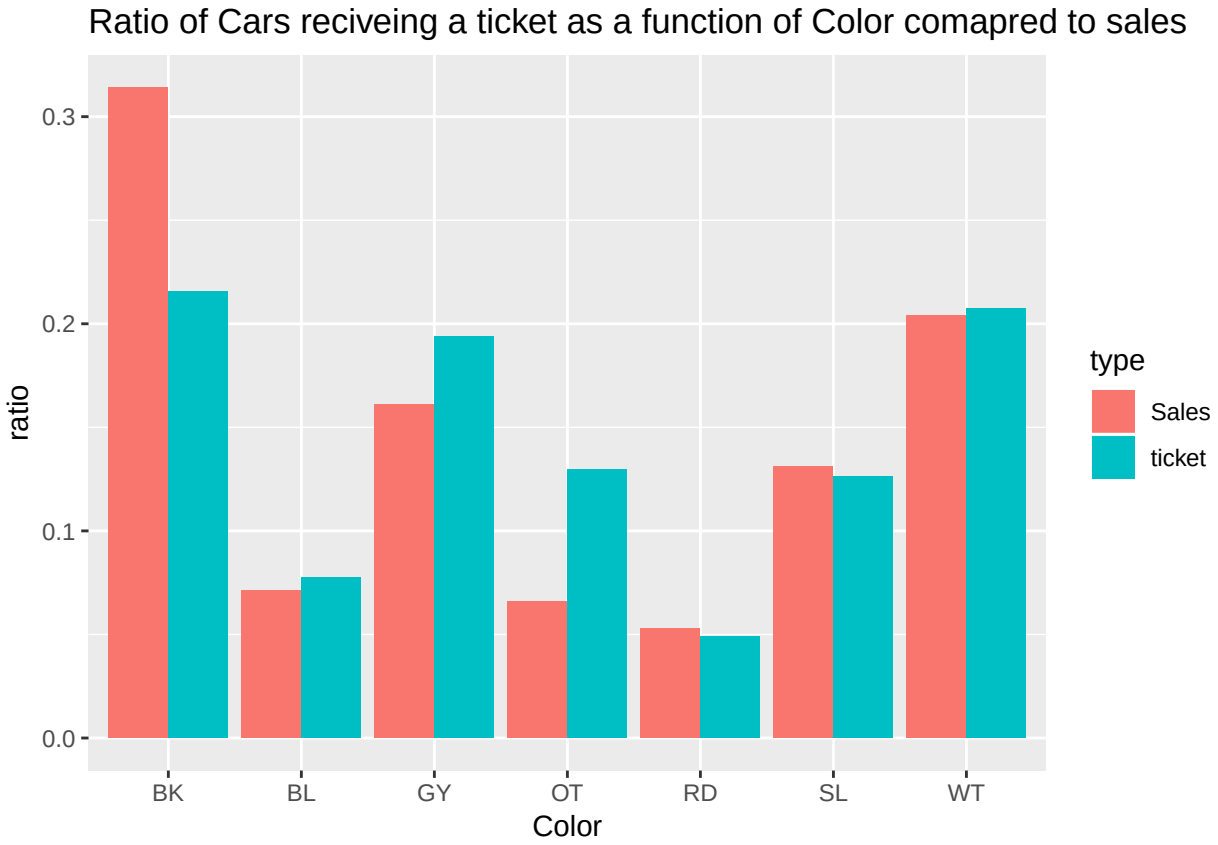
```r
##linear model  of  the female  appearance ratio by year in The Marvel universe
DC_ratio_fit<-lm(DC_FM_ratio_df$ratio~DC_FM_ratio_df$year)
summary(DC_ratio_fit)
```

```
##
## Call:
## lm(formula = DC_FM_ratio_df$ratio ~ DC_FM_ratio_df$year)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.037834 -0.010856  0.004561  0.011131  0.044041
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.538e+00  1.730e-01   8.887 3.04e-13 ***
## DC_FM_ratio_df$year -5.646e-04  8.769e-05  -6.439 1.12e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01644 on 73 degrees of freedom
## Multiple R-squared:  0.3623, Adjusted R-squared:  0.3535
## F-statistic: 41.47 on 1 and 73 DF,  p-value: 1.12e-08
```

Tickets given as a function of color Code:

```
smalldat<-read.csv(here('Final Project','parking-sub-dat.csv'))
SaleCarCol<-read.csv(here('Final Project','CarColormod.csv'))
SaleCarCol$type<-"Sales"
smalldat$Color[smalldat$Color!="BK"&smalldat$Color!="BL"&smalldat$Color!="GY"&smalldat$Color!="RD"&smal

Color<-c("BK","BL","GY","RD","SL","WT","OT")
ratio<-rep(0,7)
number<-rep(0,7)
for(col in seq(1,7)){
ratio[col]<-(nrow(smalldat[smalldat$Color==Color[col],])/nrow(smalldat))
number[col]<-nrow(smalldat[smalldat$Color==Color[col],])
}
TicketCarCol<-data.frame(Color,ratio,number)
TicketCarCol$type<-"ticket"

#chisq.test(TicketCarCol$number,p=SaleCarCol$ratio)

ratio_tot<-rbind(TicketCarCol[,-3],SaleCarCol)
ggplot(data = ratio_tot, aes(x=Color,y=ratio,fill=type)) +
geom_bar(stat = "identity",position = 'dodge')+
ggtitle('Ratio of Cars reciveing a ticket as a function of Color comapred to sales ')
```

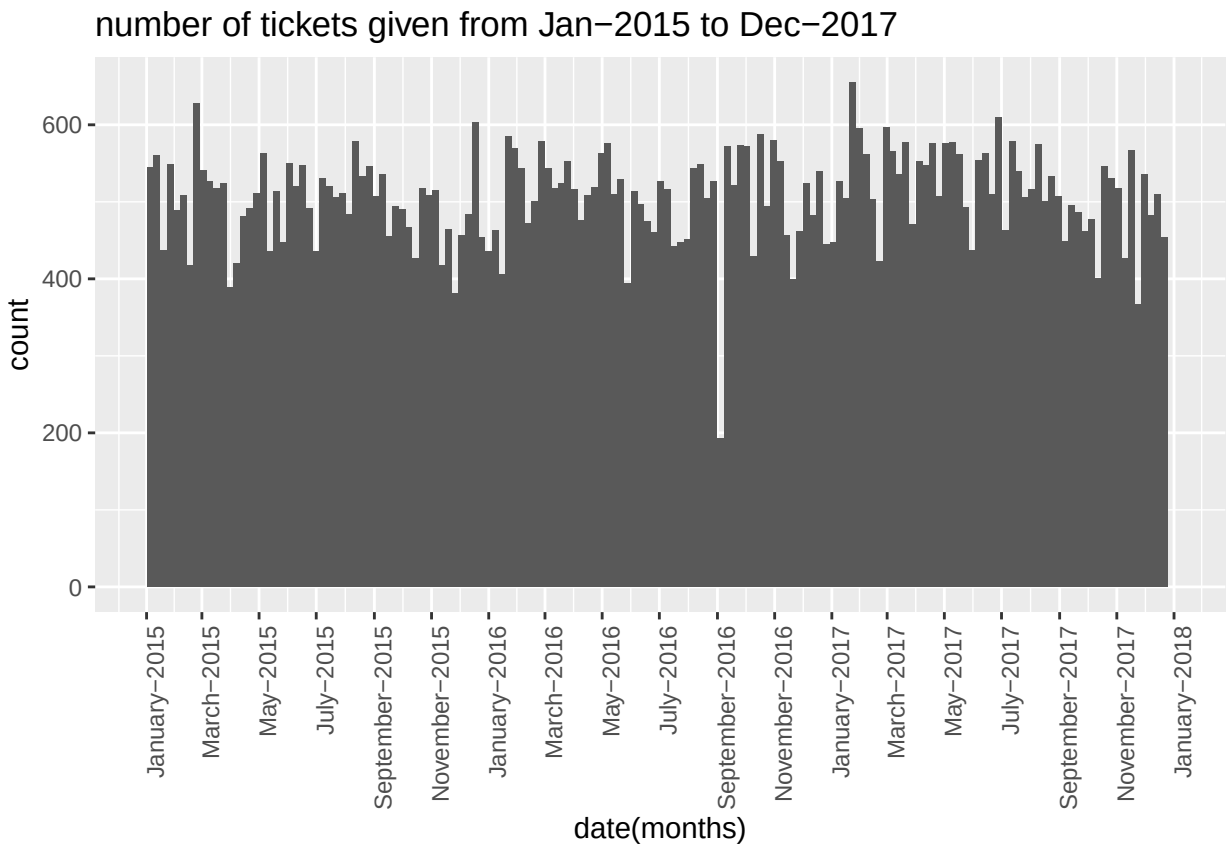


Histogram of Year, Month and Day Code:

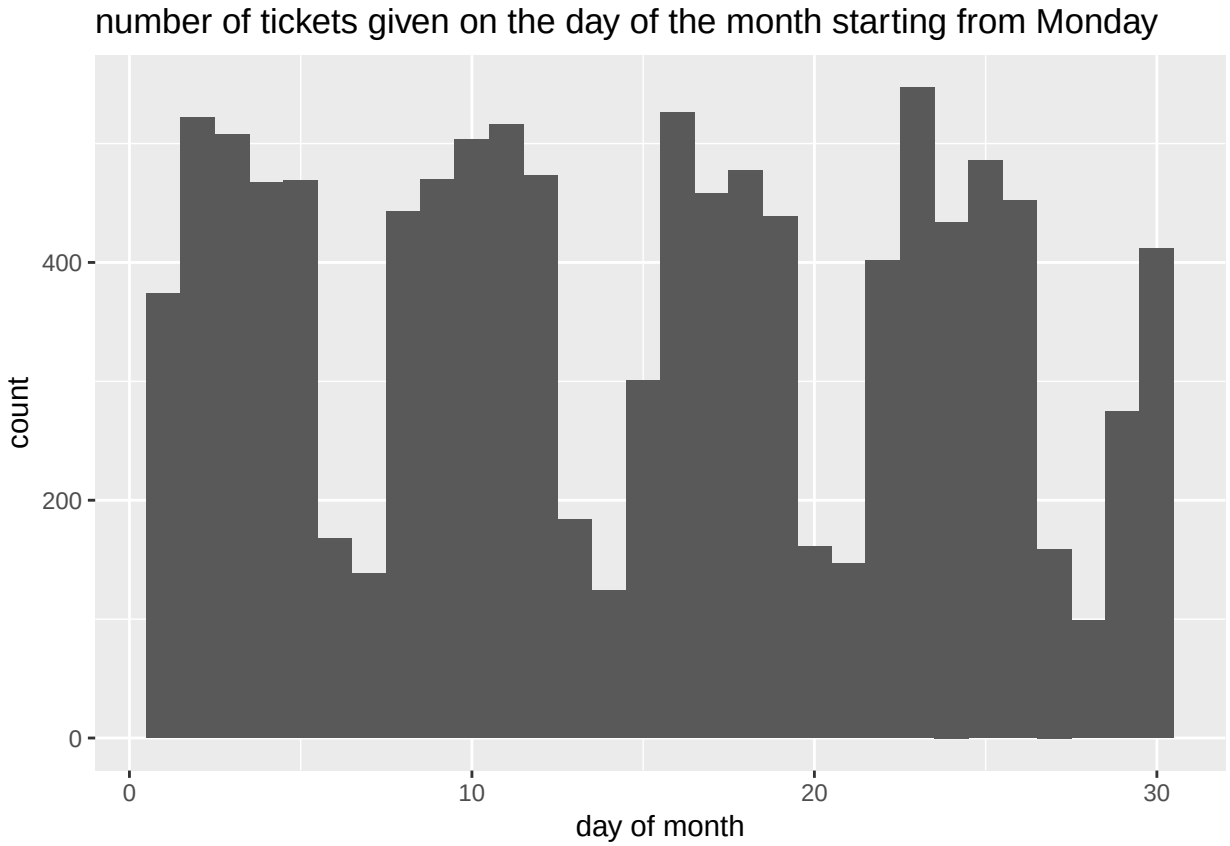
```
##----------Year-----------------

#na.omit deletes any field that hase 'NA' as value. in this case the value "date" only
#keeps the fields in Issue.Date of the subsample that have a value and are not 'NA'.
date<-na.omit(smalldat$Issue.Date)
#gsub substitues the characters in this case the time ('T00:00:00') with nothing ('').
#we are using this code to delete those parts form the text.
date<-gsub('T00:00:00','',date)
# this code turns the strings that are formated year-month-day in the list into dates.
date<-as.Date(date,format="%Y-%m-%d")


ggplot(data=NULL,aes(x=date))+
geom_histogram(bins=52*3)+
scale_x_date(labels=date_format("%B-%Y"),breaks = "2 months",limits=c(as.Date("2015-01-1"),as.Date("2018
# to see more date formats check the link below
# https://www.statmethods.net/input/dates.html
# to change the range change the date values for examle to have all of 2017 you can use:
#scale_x_date(labels=date_format("%b"),breaks = "1 month",limits=c(as.Date("2017-01-1"),as.Date("2018-0
xlab('date(months)')+
ggtitle('number of tickets given from Jan-2015 to Dec-2017')+
theme(axis.text.x=element_text(angle=90,hjust=1))
```

```
## Warning: Removed 21555 rows containing non-finite values (stat_bin).
```

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

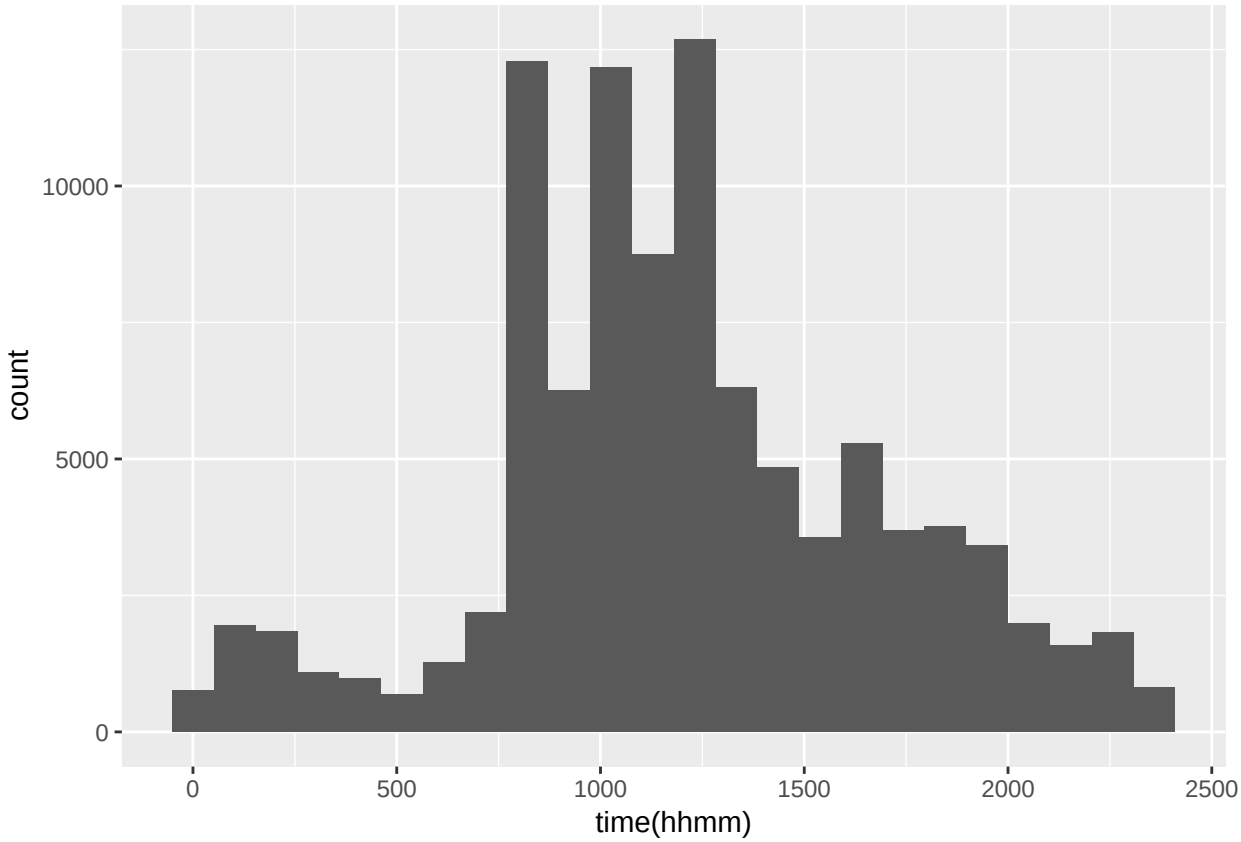



```
##----------Month--------------------

## months begining with Moday: https://www.timeanddate.com/calendar/weekday-monday-1
date<-sort(date)
begin<-c('2015-06-01','2016-02-01','2016-08-01','2017-05-01','2018-01-01')
end<-c('2015-06-30','2016-02-28','2016-08-30','2017-05-30','2018-01-30')
firstmonday<-NULL
for (i in seq(5) ){
    firstmonday<-append(firstmonday,date[date >= as.Date(begin[i]) & date <= as.Date(end[i])])
}
firstmonday_Day<-as.integer(format(firstmonday,"%d"))
ggplot(data=NULL,aes(x=firstmonday_Day))+
geom_histogram(bins=30)+
xlab('day of month')+
ggtitle('number of tickets given on the day of the month starting from Monday ')
```




```
##----------Day--------------------
ggplot(data=smalldat,aes(smalldat$Issue.time))+
geom_histogram(bins=24)+
xlab('time(hhmm)')
```

```
## Warning: Removed 25 rows containing non-finite values (stat_bin).
```

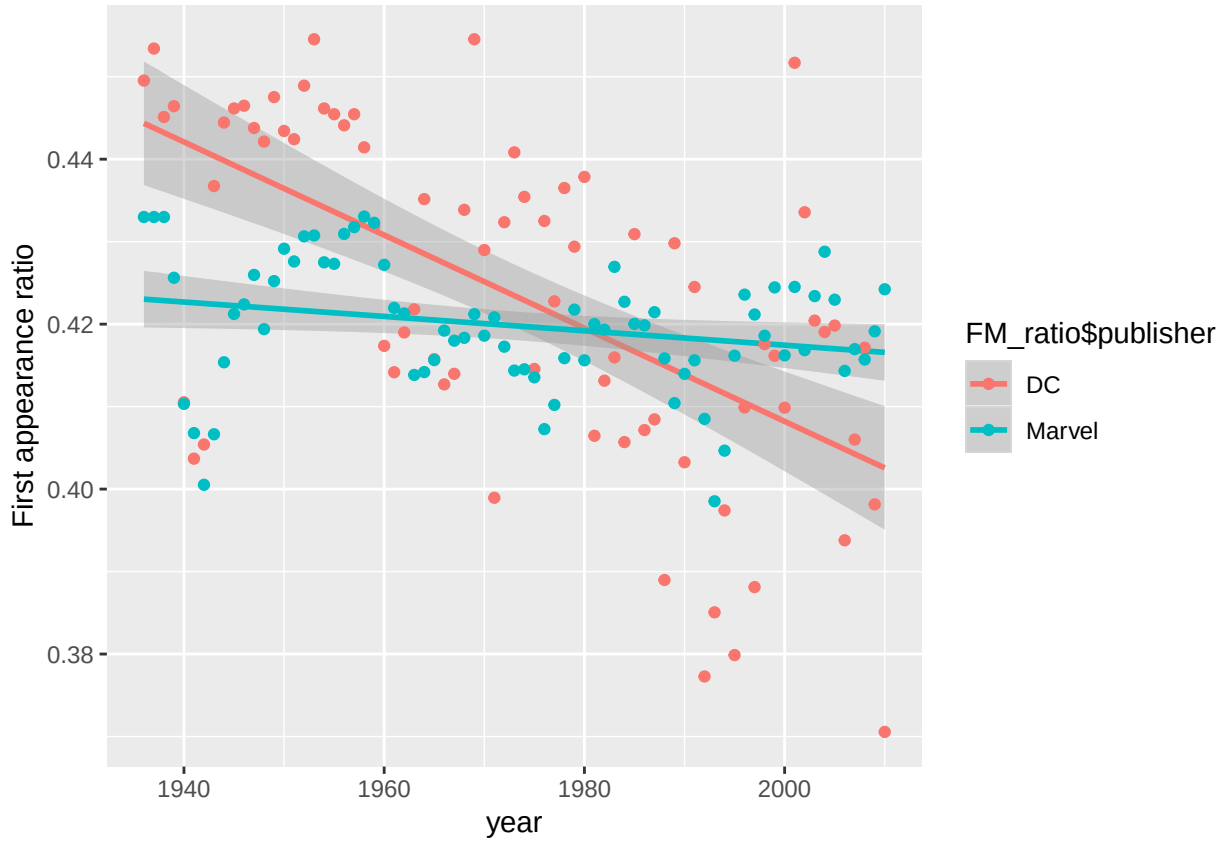



```
#lm(smalldat$Color~dat2$Make)
#dat$Issue.time
```

Ratio of female Comic Characters over years in DC and Marvel:

```
#This the type of coding I hate, I would have rather done this using dataframes but we were in a hurry.

datcomic<-read.csv(here('Final Project','tidytuesday-comic.csv'))
datcomic$gsm <-NULL
DC_comic<-datcomic[datcomic$publisher=='DC',]
Marvel_comic<-datcomic[datcomic$publisher=='Marvel',]

DC_Fem_comic<-DC_comic[DC_comic$sex == 'Female Characters',]
DC_Male_comic<-DC_comic[DC_comic$sex == 'Male Characters',]

Marvel_Fem_comic<-Marvel_comic[Marvel_comic$sex == 'Female Characters',]
Marvel_Male_comic<-Marvel_comic[Marvel_comic$sex == 'Male Characters',]
DC_FM_ratio<-rep(0,90)
DC_app_ratio<-rep(0,90)
Marvel_FM_ratio<-rep(0,90)
Marvel_app_ratio<-rep(0,90)
FM_ratio<-data.frame(matrix(ncol = 2, nrow = 150))
for (I in seq(1936,2010,by=1)) {

    Fem_year<-nrow(DC_Fem_comic[DC_Fem_comic$year==I,])
    Fem_app<-sum(na.omit(DC_Fem_comic[DC_Fem_comic$year==I,'appearances']))
```

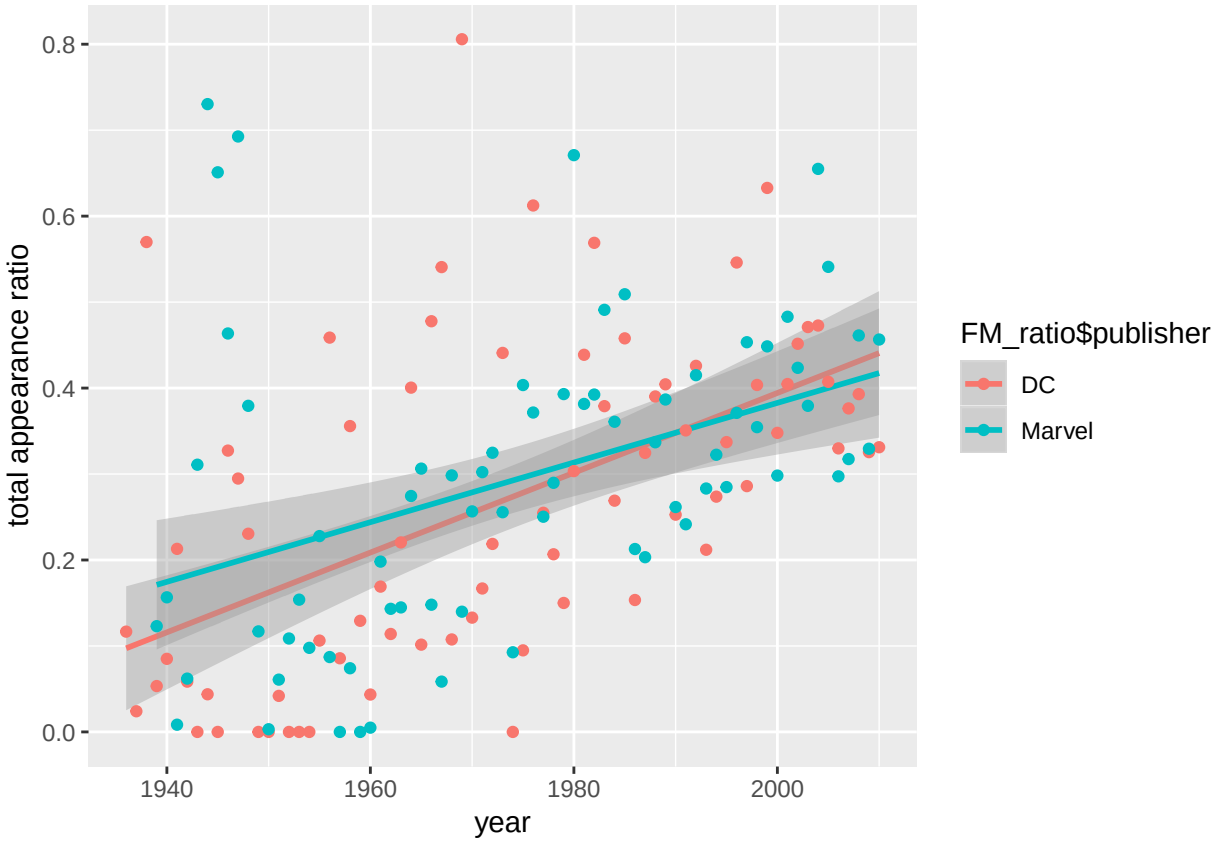
```
    Male_year<-nrow(DC_Male_comic[DC_Male_comic$year==I,])
    Male_app<-sum(na.omit(DC_Male_comic[DC_Male_comic$year==I,'appearances']))
    DC_app_ratio[I]<-(Fem_app/(Male_app+Fem_app))
    DC_FM_ratio[I]<-(Fem_year/(Male_year+Fem_year))

}
for (I in seq(1936,2010,by=1)){
    Fem_year<-nrow(Marvel_Fem_comic[Marvel_Fem_comic$year==I,])
    Fem_app<-sum(na.omit(Marvel_Fem_comic[Marvel_Fem_comic$year==I,'appearances']))
    Male_year<-nrow(Marvel_Male_comic[Marvel_Male_comic$year==I,])
    Male_app<-sum(na.omit(Marvel_Male_comic[Marvel_Male_comic$year==I,'appearances']))
    Marvel_FM_ratio[I]<-Fem_year/(Male_year+Fem_year)
    Marvel_app_ratio[I]<-(Fem_app/(Male_app+Fem_app))


}
Marvel_FM_ratio_df<-data.frame(ratio=Marvel_FM_ratio[1936:2010])
Marvel_FM_ratio_df$app_ratio<-Marvel_app_ratio[1936:2010]
Marvel_FM_ratio_df$year<-seq(1936,2010,by=1)
Marvel_FM_ratio_df$publisher<-'Marvel'

DC_FM_ratio_df<-data.frame(ratio=DC_FM_ratio[1936:2010])
DC_FM_ratio_df$app_ratio<-DC_app_ratio[1936:2010]
DC_FM_ratio_df$year<-seq(1936,2010,by=1)
DC_FM_ratio_df$publisher<-'DC'

FM_ratio<-rbind(DC_FM_ratio_df,Marvel_FM_ratio_df)

ggplot(data=FM_ratio,aes(x=FM_ratio$year,y=FM_ratio$ratio,color=FM_ratio$publisher))+
stat_smooth(method="lm")+
xlab('year')+
ylab('First appearance ratio')+
geom_point()
```




```
ggplot(data=FM_ratio,aes(x=FM_ratio$year,y=FM_ratio$app_ratio,color=FM_ratio$publisher))+
stat_smooth(method="lm")+
xlab('year')+
ylab('total appearance ratio')+
geom_point()
```

```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



Chi square test for associating hair and eye color to alignment

```
datcomic<-read.csv(here('Final Project','tidytuesday-comic.csv'))
datcomic$hair[datcomic$hair=='No Hair']<-"Bald"


Chi_hair<-datcomic[datcomic$hair=='Black Hair'|datcomic$hair=='Blond Hair'|datcomic$hair=='Bald'|datcomi


chisq.test(Chi_hair$hair, Chi_hair$align)
```

```
## Warning in chisq.test(Chi_hair$hair, Chi_hair$align): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Chi_hair$hair and Chi_hair$align
## X-squared = 487.21, df = 12, p-value < 2.2e-16
```

```
table(Chi_hair$hair, Chi_hair$align)
```

```
##
##                       Bad Characters Good Characters Reformed Criminals
##   Auburn Hair                      0               0                  0
##   Bald                          1137             329                  0
##   Black Hair                    2205            1863                  0
##   Blond Hair                     719             980                  2
```

```
##   Blue Hair                          0              0                 0
##   Bronze Hair                        0              0                 0
##   Brown Hair                      1257           1315                 0
##   Dyed Hair                          0              0                 0
##   Gold Hair                          0              0                 0
##   Green Hair                         0              0                 0
##   Grey Hair                        262            227                 0
##   Light Brown Hair                   0              0                 0
##   Magenta Hair                       0              0                 0
##   No Hair                            0              0                 0
##   Orange Hair                        0              0                 0
##   Orange-brown Hair                  0              0                 0
##   Pink Hair                          0              0                 0
##   Platinum Blond Hair                0              0                 0
##   Purple Hair                        0              0                 0
##   Red Hair                         345            433                 1
##   Reddish Blond Hair                 0              0                 0
##   Reddish Brown Hair                 0              0                 0
##   Silver Hair                        0              0                 0
##   Strawberry Blond Hair              0              0                 0
##   Variable Hair                      0              0                 0
##   Violet Hair                        0              0                 0
##   White Hair                       450            404                 0
##   Yellow Hair                        0              0                 0
```

```
Chi_eye<-datcomic[datcomic$eye=='Black Eyes'|datcomic$eye=='Blue Eyes'|datcomic$eye=='Brown Eyes'|datcor

chisq.test(Chi_eye$eye, Chi_eye$align)
```

```
## Warning in chisq.test(Chi_eye$eye, Chi_eye$align): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Chi_eye$eye and Chi_eye$align
## X-squared = 386.1, df = 16, p-value < 2.2e-16
```

```
table(Chi_eye$hair, Chi_eye$align)
```

```
##
##                       Bad Characters Good Characters Reformed Criminals
##   Auburn Hair                     12              26                 0
##   Bald                           594             197                 0
##   Black Hair                     965            1126                 0
##   Blond Hair                     384             604                 1
##   Blue Hair                       27              16                 0
##   Bronze Hair                      0               0                 0
##   Brown Hair                     601             734                 0
##   Dyed Hair                        0               0                 0
##   Gold Hair                        1               4                 0
##   Green Hair                      45              45                 0
##   Grey Hair                      119             104                 0
##   Light Brown Hair                 0               2                 0
##   Magenta Hair                     0               1                 0
##   No Hair                          0               0                 0
```

```
##   Orange Hair                5              14               0
##   Orange-brown Hair          0               0               0
##   Pink Hair                  8               9               0
##   Platinum Blond Hair        1               0               0
##   Purple Hair                8              16               0
##   Red Hair                 177             254               1
##   Reddish Blond Hair         2               3               0
##   Reddish Brown Hair         1               2               0
##   Silver Hair                4               6               0
##   Strawberry Blond Hair      9              29               0
##   Variable Hair              1               2               0
##   Violet Hair                1               0               0
##   White Hair               200             209               0
##   Yellow Hair                5               3               0
```