Analyse_5: Predictive models.

Georgie Knight

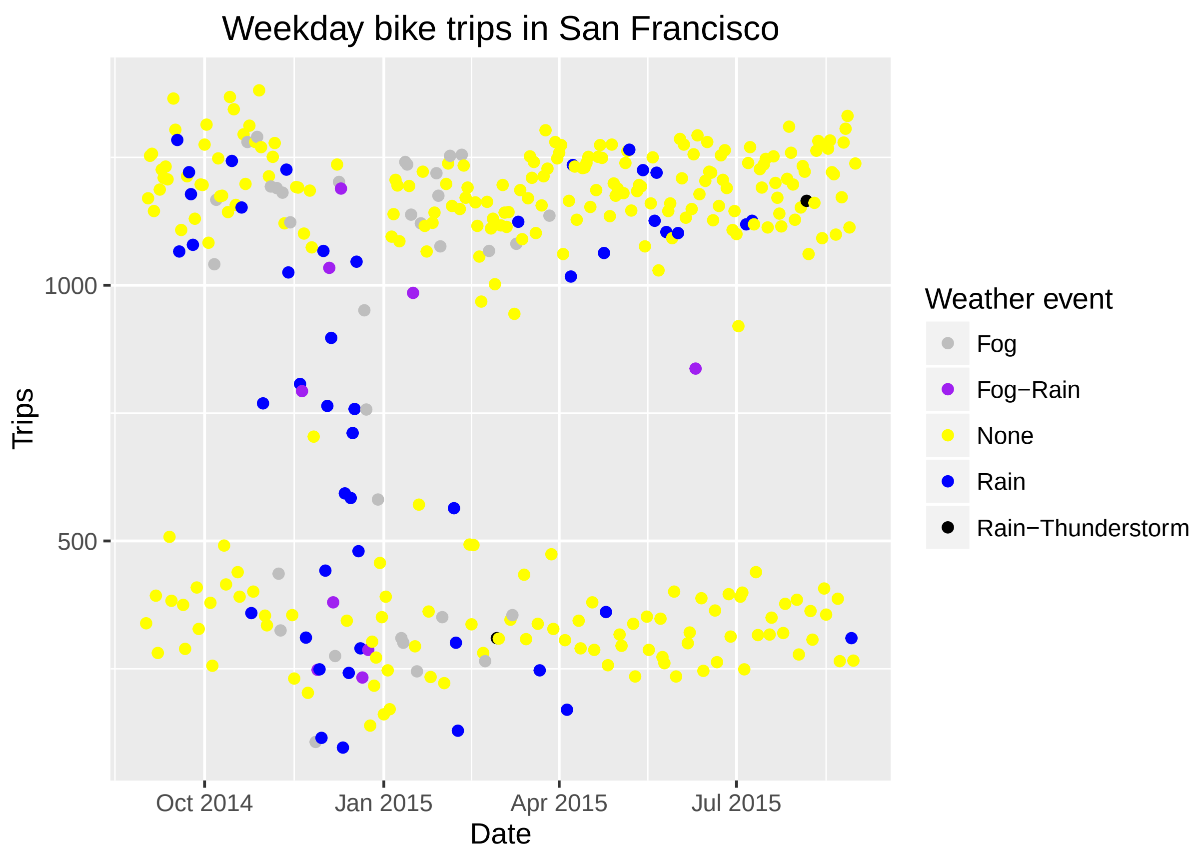
20 September, 2016

```
library("dplyr")
library("tidyr")
library("lubridate")
library("readr")
library("ggplot2")
library("ggmap")
library("igraph")
library("popgraph")
library("zipcode")
trip_read    <- read_csv("trip_full_updated2.csv")
status_read  <- read_csv("status_full_updated.csv")
weather_read <- read_csv("201508_weather_data.csv")
trip         <- dplyr::tbl_df(trip_read)
status       <- dplyr::tbl_df(status_read)
weather       <- dplyr::tbl_df(weather_read)
```

Introduction

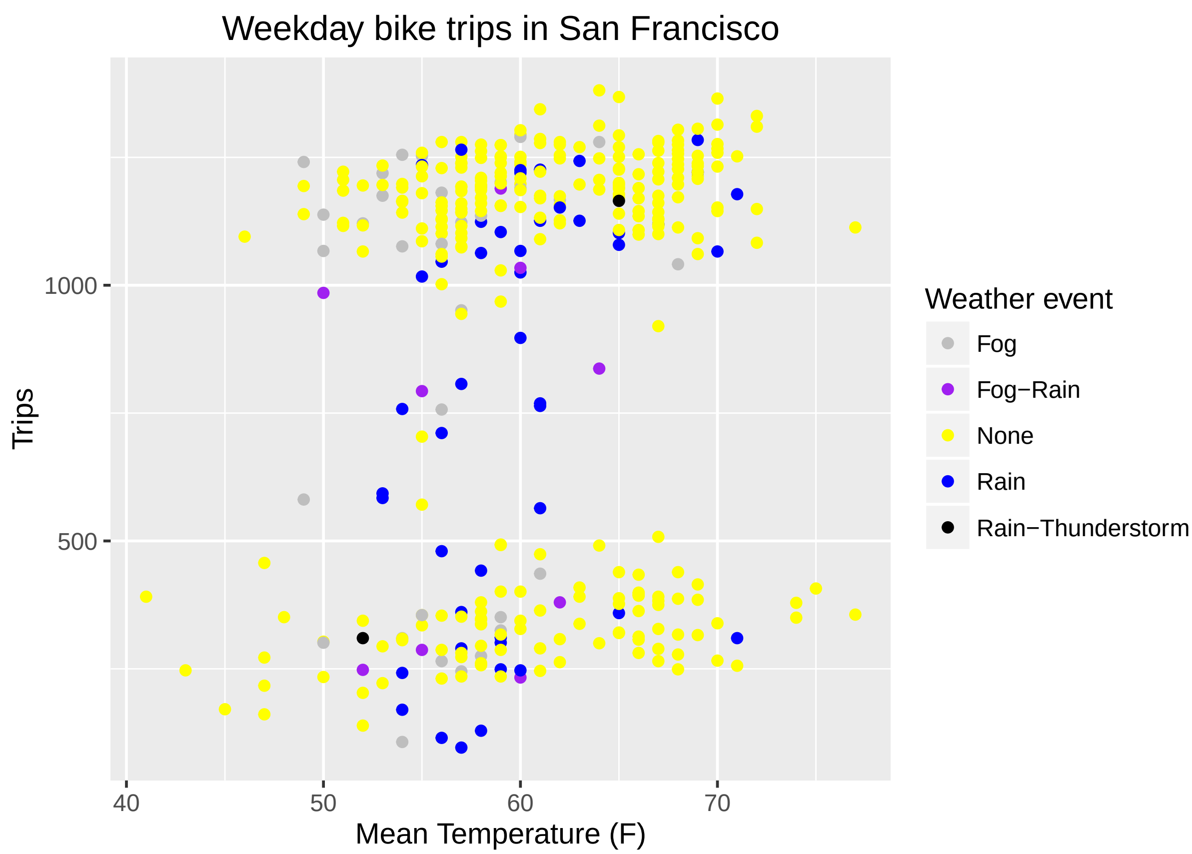
Let's try to create some predictive models. Let's first take a look at the trips per day during the week as we know that usage drops 70% on the weekends. Furthermore we'll concentrate on San Francisco.

```
trip_by_day <- trip  %>%
  filter(startLandmark == "San Francisco",
         endLandmark == "San Francisco")  %>%
  group_by(Date, Weekday, Events, Mean.TemperatureF)  %>%
  summarise(count = n()) %>%
  ungroup() %>%
  replace_na(list(Events = "None")) %>%
  mutate(Events = as.factor(Events))

ggplot(trip_by_day, aes(x= Date, y = count, col = Events))+
   geom_point()+
   labs(x = "Date", y = "Trips", col = "Weather event",
        title = "Weekday bike trips in San Francisco")+
   scale_color_manual(values=c("grey", "purple" ,"yellow", "blue", "black"))
```



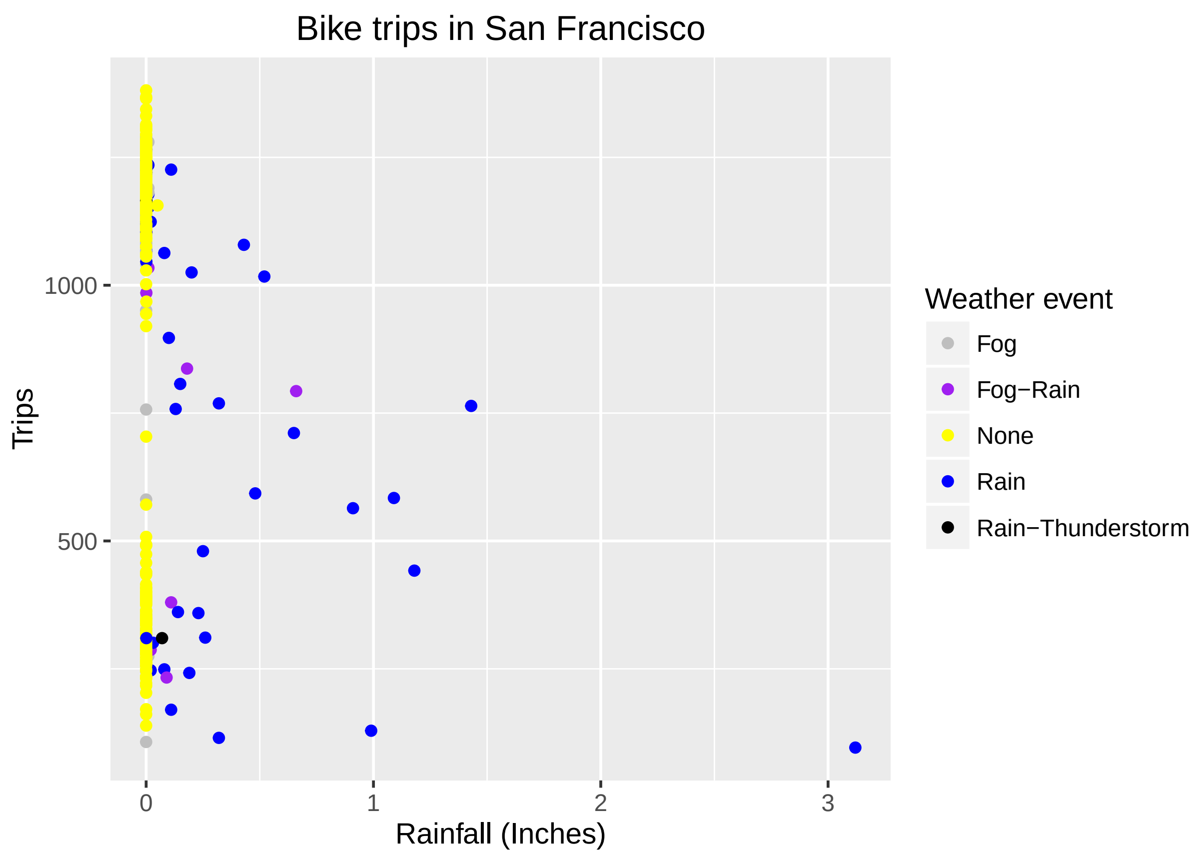
We can look at it in terms of mean temperature

```
ggplot(trip_by_day, aes(x= Mean.TemperatureF, y = count, col = Events))+
     geom_point()+
     labs(x = "Mean Temperature (F)", y = "Trips", col = "Weather event",
          title = "Weekday bike trips in San Francisco")+
     scale_color_manual(values=c("grey", "purple" ,"yellow", "blue", "black"))
```

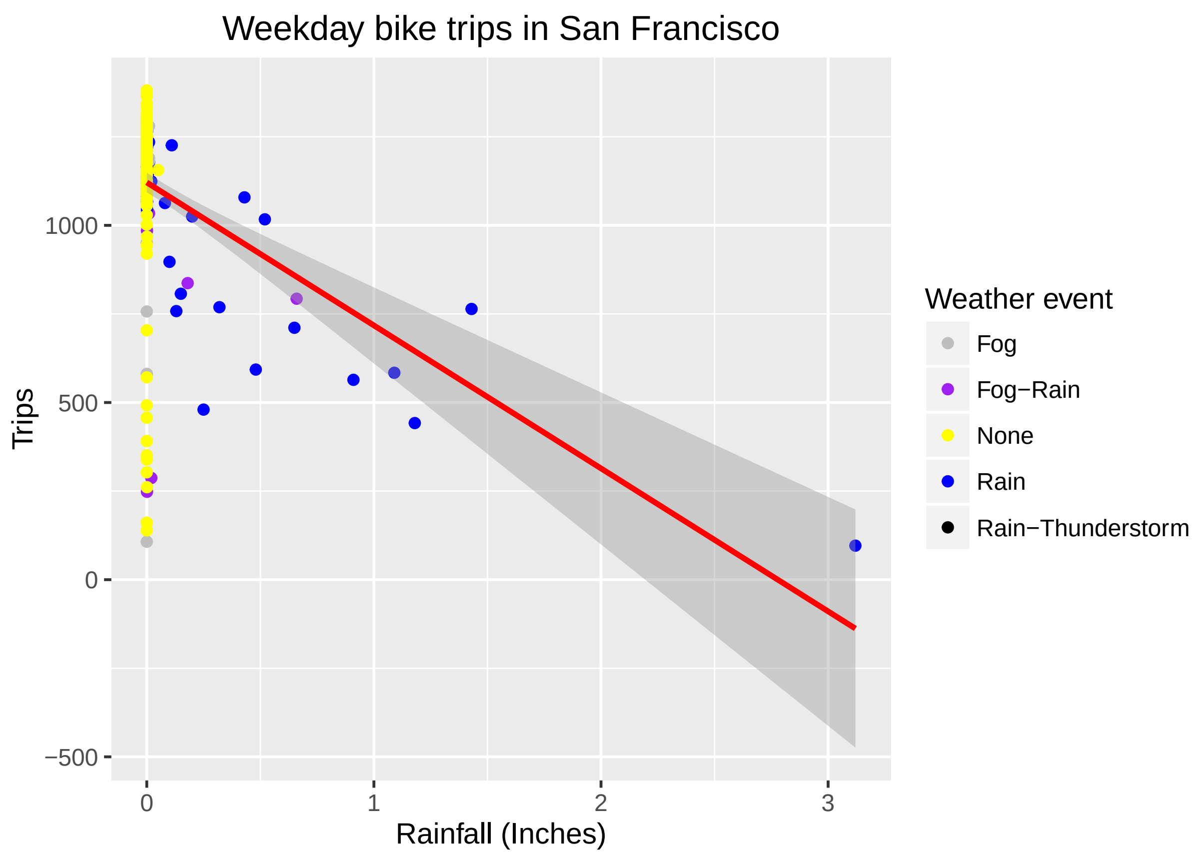


We see the general drop in usage around January, also some intersting outliers. These could be holidays. We perhaps need more specific weather information to analyse its affect on bike usage.

```
weatherSF <- weather %>%
  filter(Zip == 94107) %>%
  mutate( Date = as.Date(PDT, "%m/%d/%Y")) %>%
  select(Date, `Mean TemperatureF`, Events, PrecipitationIn) %>%
  mutate(PrecipitationIn = as.numeric(PrecipitationIn)) %>%
  replace_na(list(Events = "None", PrecipitationIn = 0.001)) %>%
  mutate(Date = ymd(Date))

trip_by_day$Rain <-weatherSF$PrecipitationIn

ggplot(trip_by_day, aes(x= Rain, y = count, col = Events))+
     geom_point()+
     labs(x = "Rainfall (Inches)", y = "Trips", col = "Weather event",
          title = "Bike trips in San Francisco")+
     scale_color_manual(values=c("grey", "purple" ,"yellow", "blue", "black"))
```




```r
trip_weekday <- filter(trip_by_day, Weekday != "Saturday" & Weekday != "Sunday")

ggplot(trip_weekday, aes(x= Rain, y = count, col = Events))+
     geom_point()+
     labs(x = "Rainfall (Inches)", y = "Trips", col = "Weather event",
          title = "Weekday bike trips in San Francisco")+
     scale_color_manual(values=c("grey", "purple" ,"yellow", "blue", "black"))+
     stat_smooth(method = "lm", col = "red")
```




```r
trip_weekday %>% group_by(Events) %>%
  summarise(mean_trips = mean(count))
```

```
## # A tibble: 5 x 2
##              Events mean_trips
##              <fctr>      <dbl>
## 1               Fog  1082.4400
## 2          Fog-Rain   767.5714
## 3              None  1141.5026
## 4              Rain   965.0857
## 5 Rain-Thunderstorm  1165.0000
```

```r
filter(trip_weekday, Events == "Rain-Thunderstorm")
```

```
## # A tibble: 1 x 6
##         Date  Weekday            Events Mean.TemperatureF count  Rain
##       <time>    <chr>            <fctr>             <int> <int> <dbl>
## 1 2015-08-06 Thursday Rain-Thunderstorm                65  1165 0.001
```

Let's look at this in terms of rained or not:

```r
trip_weekday %>% mutate(Rained = ifelse(Rain >0.01, "Rain","No Rain")) %>%
  group_by(Rained) %>% summarise(mean_trips = mean(count))
```

```
## # A tibble: 2 x 2
##    Rained mean_trips
##     <chr>      <dbl>
## 1 No Rain    1132.28
## 2    Rain     776.00
```

So when there is less than 0.01 inches of rain there are an average of 1132.28 trips per day, whilst when it rains there are 776, a drop of over 30 per cent. Is the mean of our 22 observations out of 261 statistically significant? We'll calculate the probability that a mean of 776 or less is found from a random sample of 22 observations

```r
set.seed(100)
samples < 10000000
sum <- 0
for (i in 1:samples) {
  sampleMean <- mean(sample_n(trip_weekday, 22)$count)
  if(sampleMean <= 776){
    sum <- sum + 1
    }
}
print(sum/samples)
```

Running the above code gives a p-value of around 0.0000001, highly significant.

Let's look at some models

```r
linear_model <- lm(count ~ Rain + Mean.TemperatureF, data = trip_weekday)
summary(linear_model)
```

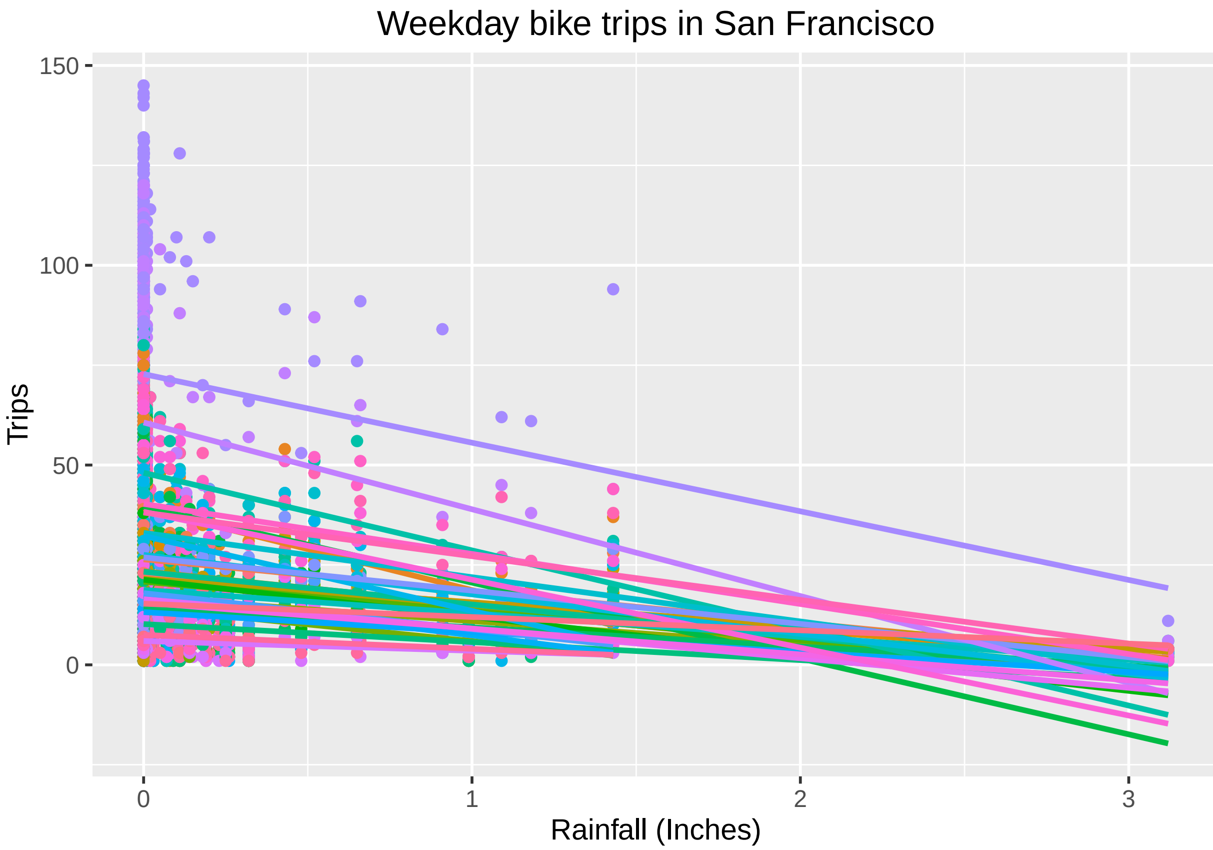
```
##
## Call:
## lm(formula = count ~ Rain + Mean.TemperatureF, data = trip_weekday)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -928.95  -23.97   46.30  117.84  270.39
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        330.288    129.714   2.546   0.0115 *
## Rain              -377.203     51.983  -7.256 4.66e-12 ***
## Mean.TemperatureF   13.068      2.132   6.129 3.30e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210.3 on 258 degrees of freedom
## Multiple R-squared:  0.2757, Adjusted R-squared:  0.2701
## F-statistic: 49.11 on 2 and 258 DF,  p-value: < 2.2e-16
```

```r
exponential_model <- lm(log(count) ~ Rain + Mean.TemperatureF, data = trip_weekday)
summary(exponential_model)
```

```
##
```

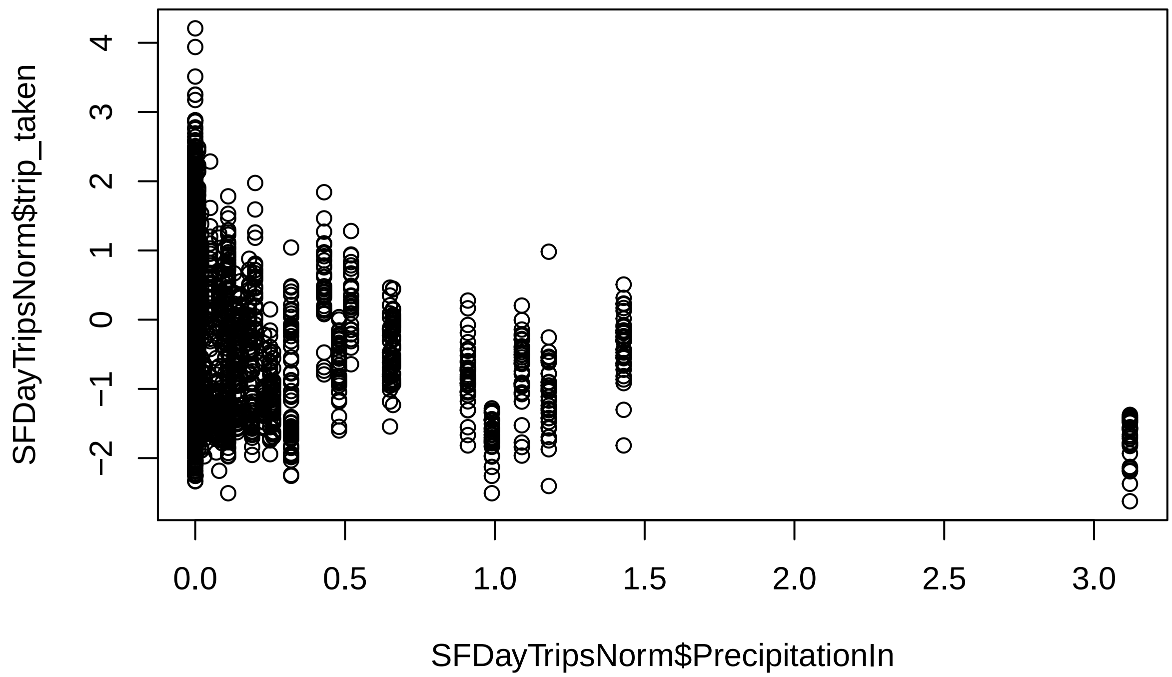
```
## Call:
## lm(formula = log(count) ~ Rain + Mean.TemperatureF, data = trip_weekday)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.19488 -0.02952  0.06940  0.15437  0.56912
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.882927   0.206379  28.506  < 2e-16 ***
## Rain              -0.647490   0.082706  -7.829 1.28e-13 ***
## Mean.TemperatureF  0.018237   0.003393   5.376 1.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3345 on 258 degrees of freedom
## Multiple R-squared:  0.2749, Adjusted R-squared:  0.2693
## F-statistic:  48.9 on 2 and 258 DF,  p-value: < 2.2e-16
```

```r
SFDayTrips <- trip  %>% filter(Weekday != "saturday", Weekday != "sunday", startLandmark =="San Francisc
  group_by(Date, Start.Station)  %>%
  summarise(count =n()) %>%
  ungroup() %>%
  mutate(Date = as.Date(Date)) %>%
  left_join(weatherSF, by =c("Date" = "Date"))

ggplot(SFDayTrips, aes(x= PrecipitationIn, y = count, col = as.factor(Start.Station)))+
   geom_point()+
   labs(x = "Rainfall (Inches)", y = "Trips", col = "Weather event",
          title = "Weekday bike trips in San Francisco")+
     stat_smooth(method = "lm", se =FALSE) + theme(legend.position="none")
```



Which gives us a linear model for each station, we can normalise the number of trips and at each stationlook at the trend

```
SFDayTripsNorm <- SFDayTrips %>%
  group_by(Start.Station) %>%
  mutate(trip_taken = scale(count)) %>%
  ungroup()
plot(y=SFDayTripsNorm$trip_taken, x=SFDayTripsNorm$PrecipitationIn)
```


SFDayTripsNorm$trip_taken
SFDayTripsNorm$PrecipitationIn
4
3
2
1
0
−1
−2
0.0
0.5
1.0
1.5
2.0
2.5
3.0