stats101c_hw4

Takao

10/31/2022

TAKAO OBA

Stats 101C HW4

Q1 Download the training and the testing data sets

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
# library(tidyverse)
acc.test <- read.csv("/Users/takaooba/Downloads/predicting-car-accidents-severity/AcctestNoYNew.csv")

acc.train <- read.csv("/Users/takaooba/Downloads/predicting-car-accidents-severity/Acctrain.csv")

acc.test <- acc.test[,-1]
```

(a) Report the dimensions of both the training and the testing data sets.

The dimensions can be found by the following

```
dim(acc.test)
```

```
## [1] 15000    43
```

```
dim(acc.train)
```

```
## [1] 35000    44
```

(b) How many numerical predictors does your data have? List them.

```
head(acc.test)
```

```
##             Start_Time             End_Time Start_Lat  Start_Lng  End_Lat
## 1 2020-01-21T17:44:00Z 2020-01-21T18:58:56Z  34.05591 -117.39639 34.05591
## 2 2021-12-25T23:33:00Z 2021-12-26T00:52:03Z  36.09308 -120.11828 36.09224
## 3 2021-12-23T06:13:00Z 2021-12-23T07:37:39Z  34.03735 -118.16990 34.04512
## 4 2020-11-03T00:24:00Z 2020-11-03T02:30:00Z  36.21964  -77.10602 36.22946
## 5 2019-03-25T11:22:15Z 2019-03-25T15:22:15Z  45.17533 -118.78111 45.16698
## 6 2017-01-14T10:45:45Z 2017-01-14T16:45:45Z  29.81356  -95.39551 29.81297
##      End_Lng Distance.mi.
## 1 -117.39639        0.000
## 2 -120.11751        0.073
## 3 -118.16957        0.537
## 4  -77.09824        0.805
## 5 -118.79216        0.789
## 6  -95.40754        0.722
##                                                                 Description
## 1                                    At I-10/San Bernardino Fwy - Accident.
## 2  Incident on I-5 SB near COALINGA AVENAL REST AREA Right shoulder closed.
## 3        Incident on I-710 NB near CESAR CHAVEZ AVE Right shoulder closed.
## 4 Incident on NC-11 NB near BRICKMILL RD Road closed. Take alternate route.
## 5                                                 At Camas St - Accident.
## 6                                        At Yale St/Shepherd Dr - Accident.
##                     Street Side        City          County State    Zipcode
## 1            Santa Ana Ave    R Bloomington San Bernardino    CA 92316-2635
## 2                    I-5 S    R       Huron          Fresno    CA      93234
## 3                  I-710 N    R Los Angeles    Los Angeles    CA      90022
## 4 NC Highway 11 Business N    L    Aulander          Bertie    NC      27805
## 5        Ukiah-Hilgard Hwy    R  Pilot Rock        Umatilla    OR      97868
## 6                  I-610 W    R     Houston          Harris    TX      77008
##   Country   Timezone Airport_Code    Weather_Timestamp Temperature.F.
## 1      US US/Pacific         KRAL 2020-01-21T17:53:00Z           56.0
## 2      US US/Pacific         KNLC 2021-12-25T23:56:00Z           43.0
## 3      US US/Pacific         KCQT 2021-12-23T05:52:00Z           56.0
## 4      US US/Eastern         KASJ 2020-11-03T00:15:00Z           41.0
## 5      US US/Pacific         KPDT 2019-03-25T10:53:00Z           55.9
## 6      US US/Central         KMCJ 2017-01-14T10:35:00Z           66.2
##   Wind_Chill.F. Humidity... Pressure.in. Visibility.mi. Wind_Direction
## 1            56          75        29.27           10.0            WNW
## 2            43          93        29.65           10.0            WSW
## 3            56          72        29.81            6.0           CALM
## 4            38          50        30.11           10.0              W
## 5            NA          42        30.00           10.0             SE
## 6            NA         100        30.35            0.5           East
##   Wind_Speed.mph. Weather_Condition Amenity  Bump Crossing Give_Way Junction
## 1             5.0            Cloudy   FALSE FALSE    FALSE    FALSE    FALSE
## 2             3.0              Fair   FALSE FALSE    FALSE    FALSE     TRUE
## 3             0.0        Light Rain   FALSE FALSE    FALSE    FALSE    FALSE
## 4             5.0              Fair   FALSE FALSE    FALSE    FALSE    FALSE
## 5            18.4             Clear   FALSE FALSE    FALSE    FALSE    FALSE
## 6             8.1          Overcast   FALSE FALSE    FALSE    FALSE    FALSE
##   No_Exit Railway Roundabout Station  Stop Traffic_Calming Traffic_Signal
## 1   FALSE   FALSE      FALSE   FALSE FALSE           FALSE          FALSE
## 2   FALSE   FALSE      FALSE   FALSE FALSE           FALSE          FALSE
## 3   FALSE   FALSE      FALSE   FALSE FALSE           FALSE          FALSE
## 4   FALSE   FALSE      FALSE   FALSE FALSE           FALSE          FALSE
```

```
## 5   FALSE   FALSE      FALSE   FALSE FALSE          FALSE          FALSE
## 6   FALSE   FALSE      FALSE   FALSE FALSE          FALSE          FALSE
##   Turning_Loop Sunrise_Sunset Civil_Twilight Nautical_Twilight
## 1        FALSE          Night          Night               Day
## 2        FALSE          Night          Night             Night
## 3        FALSE          Night          Night               Day
## 4        FALSE          Night          Night             Night
## 5        FALSE            Day            Day               Day
## 6        FALSE            Day            Day               Day
##   Astronomical_Twilight
## 1                   Day
## 2                 Night
## 3                   Day
## 4                 Night
## 5                   Day
## 6                   Day
```

```
# colnames(acc.test)

head(acc.train)
```

```
##   Severity           Start_Time             End_Time Start_Lat  Start_Lng
## 1     MILD 2021-01-06T13:58:00Z 2021-01-06T16:41:33Z  30.30729  -97.72562
## 2     MILD 2021-04-03T11:08:00Z 2021-04-03T12:26:27Z  38.07553 -122.54181
## 3   SEVERE 2019-05-09T14:19:17Z 2019-05-09T14:46:43Z  25.81245  -80.21472
## 4     MILD 2021-11-15T12:22:30Z 2021-11-15T12:46:30Z  34.99765  -82.05707
## 5     MILD 2021-12-14T10:24:32Z 2021-12-14T12:09:32Z  45.50310 -118.42311
## 6     MILD 2020-04-11T04:34:29Z 2020-04-11T05:09:27Z  38.61153 -121.51080
##    End_Lat    End_Lng Distance.mi.
## 1 30.30702  -97.72503        0.040
## 2 38.07913 -122.54512        0.307
## 3 25.81242  -80.21219        0.157
## 4 34.99737  -82.05515        0.110
## 5 45.50258 -118.42264        0.042
## 6 38.61153 -121.51080        0.000
##                                                                                                Descrip
## 1                                                 Incident on E 45TH ST near AVENUE H Drive with cau
## 2                                                    Incident on US-101 NB near CA-37 Drive with cau
## 3                       Ramp closed to I-95 and I-95 Northbound Express Ln - Road closed due to accid
## 4 Stationary traffic on SC-40 from Mitchell Rd (New Cut Rd) to John Dodd Rd (New Cut Rd) due to accid
## 5                                                     Incident on I-84 EB near MP 238 Drive with cau
## 6                                     At Garden Hwy/Exit 521A/Exit 521 - Accident. Hard shoulder blo
##          Street Side       City      County State    Zipcode Country
## 1       Avenue H    R     Austin      Travis    TX 78751-3122      US
## 2  Redwood Hwy S    R     Novato       Marin    CA      94945      US
## 3 Airport Expy E    R      Miami  Miami-Dade    FL      33127      US
## 4     New Cut Rd    R      Inman Spartanburg    SC 29349-4532      US
## 5         I-84 E    R  Pendleton    Umatilla    OR      97801      US
## 6        CA-16 E    R Sacramento  Sacramento    CA      95833      US
##     Timezone Airport_Code    Weather_Timestamp Temperature.F. Wind_Chill.F.
## 1 US/Central         KATT 2021-01-06T13:51:00Z             65            65
## 2 US/Pacific         KDVO 2021-04-03T11:15:00Z             54            54
## 3 US/Eastern         KMIA 2019-05-09T13:53:00Z             87            87
## 4 US/Eastern         KSPA 2021-11-15T12:15:00Z             55            55
```

```
## 5 US/Pacific         KPDT 2021-12-14T10:53:00Z             38             30
## 6 US/Pacific         KMCC 2020-04-11T04:50:00Z             46             44
##   Humidity... Pressure.in. Visibility.mi. Wind_Direction Wind_Speed.mph.
## 1          50        29.17             10             NW              10
## 2          67        30.10             10            WNW               9
## 3          61        29.96             10             SE              13
## 4          38        29.29             10              W               3
## 5          60        28.20             10             SW              12
## 6          93        29.88             10            SSE               5
##   Weather_Condition Amenity  Bump Crossing Give_Way Junction No_Exit Railway
## 1     Partly Cloudy   FALSE FALSE     TRUE    FALSE    FALSE   FALSE   FALSE
## 2            Cloudy   FALSE FALSE    FALSE    FALSE     TRUE   FALSE   FALSE
## 3     Mostly Cloudy   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE
## 4              Fair   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE
## 5              Fair   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE
## 6              Fair   FALSE FALSE    FALSE    FALSE    FALSE   FALSE   FALSE
##   Roundabout Station  Stop Traffic_Calming Traffic_Signal Turning_Loop
## 1      FALSE   FALSE FALSE           FALSE          FALSE        FALSE
## 2      FALSE   FALSE FALSE           FALSE          FALSE        FALSE
## 3      FALSE   FALSE FALSE           FALSE          FALSE        FALSE
## 4      FALSE   FALSE FALSE           FALSE          FALSE        FALSE
## 5      FALSE   FALSE FALSE           FALSE          FALSE        FALSE
## 6      FALSE   FALSE FALSE           FALSE          FALSE        FALSE
##   Sunrise_Sunset Civil_Twilight Nautical_Twilight Astronomical_Twilight
## 1            Day            Day               Day                   Day
## 2            Day            Day               Day                   Day
## 3            Day            Day               Day                   Day
## 4            Day            Day               Day                   Day
## 5            Day            Day               Day                   Day
## 6          Night          Night             Night                 Night
```

```r
# colnames(acc.train)
```

The numerical predictors are Start_Lat, Start_Lng, End_Lat, End_Lng, Distance.mi., Temperature.F., Wind_Chill.F., Humidity…, Pressure.in., Visibility.mi., Wind_Speed.mph. There are a total of 11 numerical predictors. This is both for the training and testing data.

(c) How many categorical predictors does your data have? List them.

The categorical predictors are Street, Side, City, Country, State, Zipcode, Country, Timezone, Airport_Code, Wind_Direction, Weather_Condition, Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic Signal, Turning_Loop, Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight There are a total of 29 categorical predictors. This is both for the training and testing data.

(d) Report the size of missing values in both data sets (Training and Testing)

```r
# Testing Data
sum((is.na(acc.test)))
```

```
## [1] 5842
```

```r
# Training Data
sum(is.na(acc.train))
```

```
## [1] 13211
```

```r
sum((is.na(acc.test))) + sum(is.na(acc.train))
```

```
## [1] 19053
```

(e) Plot densities of your best six numerical predictors based on the response variable.

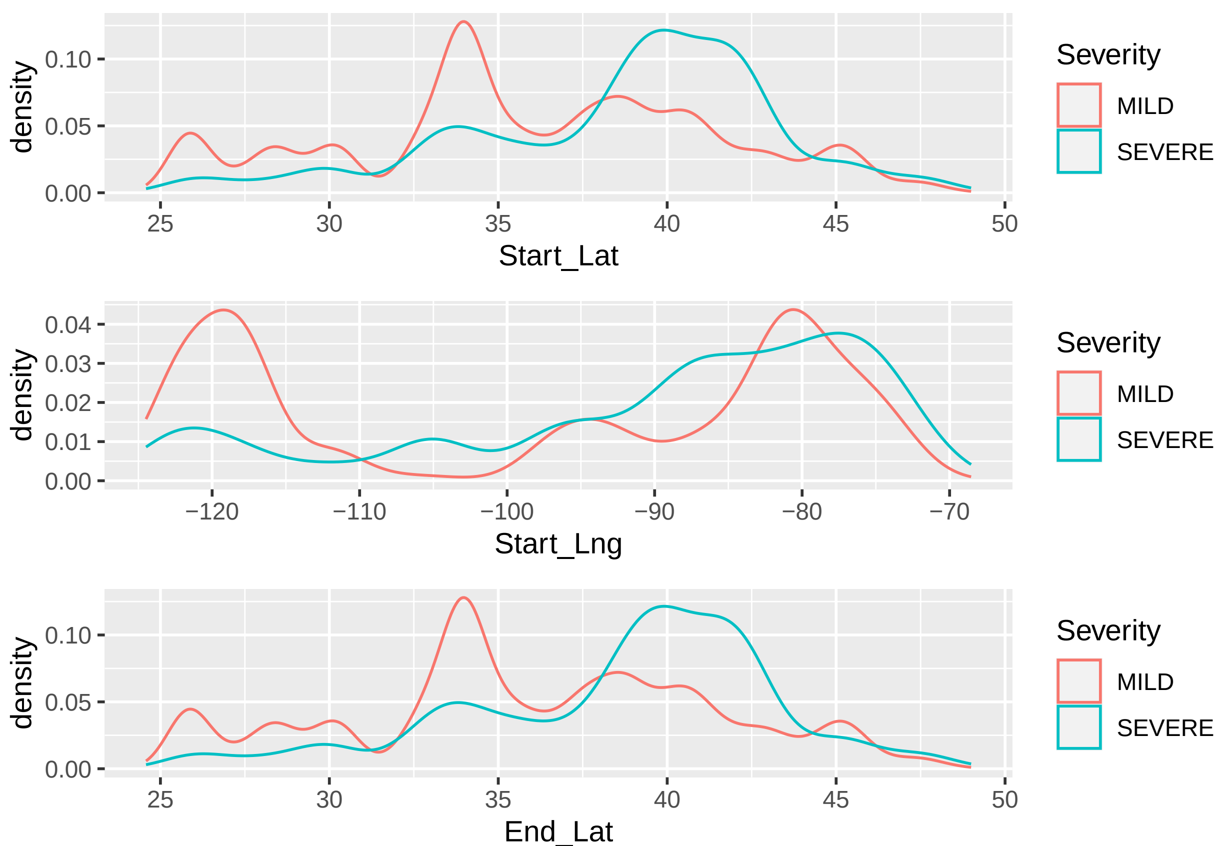
```r
acc.train.1 <- na.omit(acc.train)
# head(acc.train.1)
acc.train.1$SeverityNum <- ifelse(acc.train.1$Severity == "MILD", 0, 1)
numericalpredictor <- acc.train.1[,c(4,5,6,7,8,20,21,22,23,24,26,45)]
cor(numericalpredictor)
```

```
##                   Start_Lat   Start_Lng      End_Lat     End_Lng Distance.mi.
## Start_Lat        1.000000000 -0.16112099  0.999995357 -0.16111539   0.07759972
## Start_Lng       -0.161120989  1.00000000 -0.161127433  0.99999911   0.03032042
## End_Lat          0.999995357 -0.16112743  1.000000000 -0.16112227   0.07770666
## End_Lng         -0.161115387  0.99999911 -0.161122270  1.00000000   0.03045833
## Distance.mi.     0.077599716  0.03032042  0.077706656  0.03045833   1.00000000
## Temperature.F.  -0.493369968  0.02776738 -0.493354441  0.02776064  -0.05213097
## Wind_Chill.F.   -0.498391152  0.01064582 -0.498376917  0.01063811  -0.05731444
## Humidity...      0.007476358  0.16206716  0.007467002  0.16204723   0.02770460
## Pressure.in.    -0.261356822  0.22581863 -0.261341877  0.22580454  -0.07119964
## Visibility.mi.  -0.094135159  0.03483039 -0.094125943  0.03483449  -0.03946985
## Wind_Speed.mph.  0.033185546  0.11498774  0.033177598  0.11500546   0.02530208
## SeverityNum      0.122739891  0.10216806  0.122719408  0.10216720   0.04163956
##                 Temperature.F. Wind_Chill.F.  Humidity... Pressure.in.
## Start_Lat          -0.49336997  -0.498391152  0.007476358  -0.26135682
## Start_Lng           0.02776738   0.010645816  0.162067156   0.22581863
## End_Lat            -0.49335444  -0.498376917  0.007467002  -0.26134188
## End_Lng             0.02776064   0.010638109  0.162047225   0.22580454
## Distance.mi.       -0.05213097  -0.057314440  0.027704596  -0.07119964
## Temperature.F.      1.00000000   0.993757045 -0.374606921   0.11656740
## Wind_Chill.F.       0.99375704   1.000000000 -0.356542306   0.12255842
## Humidity...        -0.37460692  -0.356542306  1.000000000   0.15126431
## Pressure.in.        0.11656740   0.122558422  0.151264314   1.00000000
## Visibility.mi.      0.21708604   0.219076422 -0.369635702   0.02173128
## Wind_Speed.mph.     0.06196195   0.005712984 -0.170450319  -0.05529463
## SeverityNum        -0.08497974  -0.094599872  0.022847681  -0.03724430
##                 Visibility.mi. Wind_Speed.mph.  SeverityNum
## Start_Lat         -0.094135159     0.033185546  0.122739891
## Start_Lng          0.034830391     0.114987744  0.102168059
## End_Lat           -0.094125943     0.033177598  0.122719408
## End_Lng            0.034834492     0.115005458  0.102167198
## Distance.mi.      -0.039469847     0.025302077  0.041639564
```

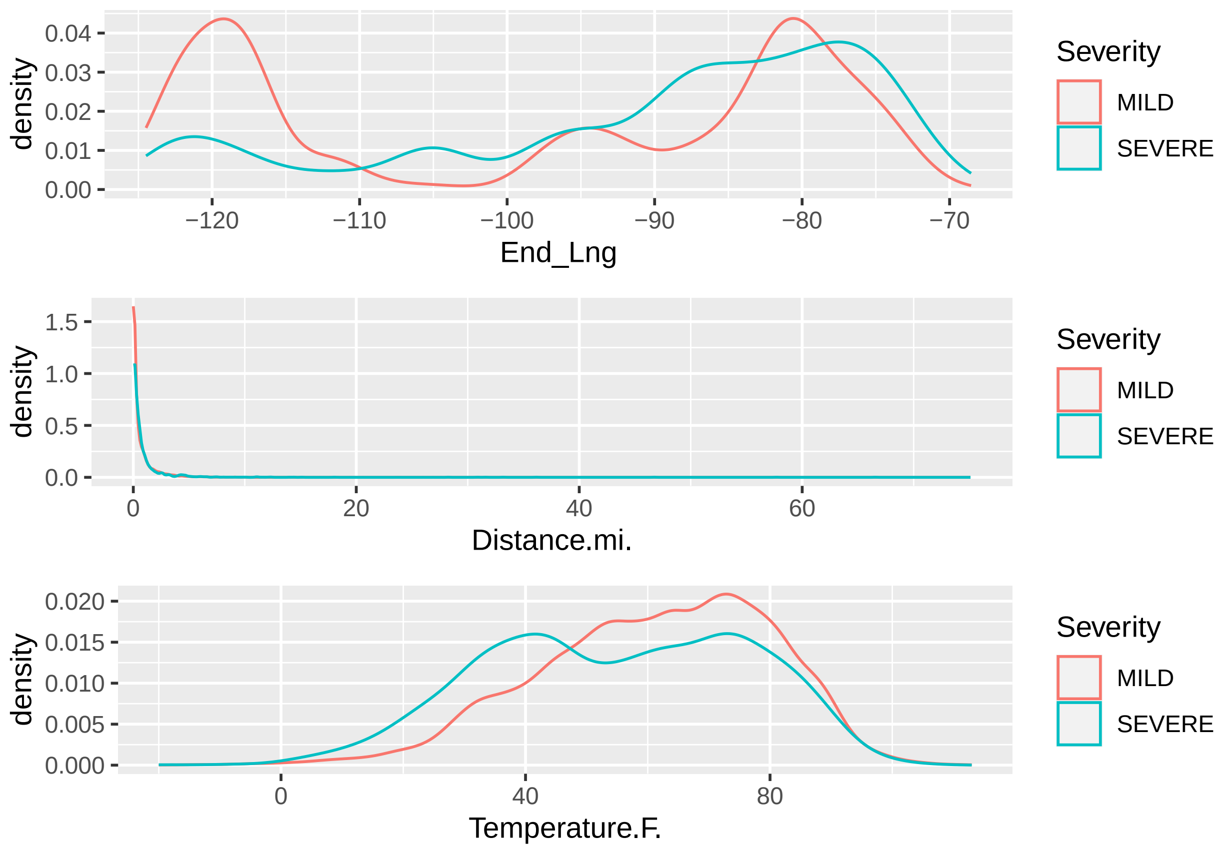
```
## Temperature.F.      0.217086037     0.061961950 -0.084979743
## Wind_Chill.F.       0.219076422     0.005712984 -0.094599872
## Humidity...        -0.369635702    -0.170450319  0.022847681
## Pressure.in.        0.021731275    -0.055294629 -0.037244301
## Visibility.mi.      1.000000000     0.025227688  0.006555872
## Wind_Speed.mph.     0.025227688     1.000000000  0.060132993
## SeverityNum         0.006555872     0.060132993  1.000000000
```

Based on the correlation plot that we have just created above, we have that the best predictors are Start_Lat, End_Lat, Start_Lng, End_Lng, Wine_Chill.F., Wind_Speed.mph.

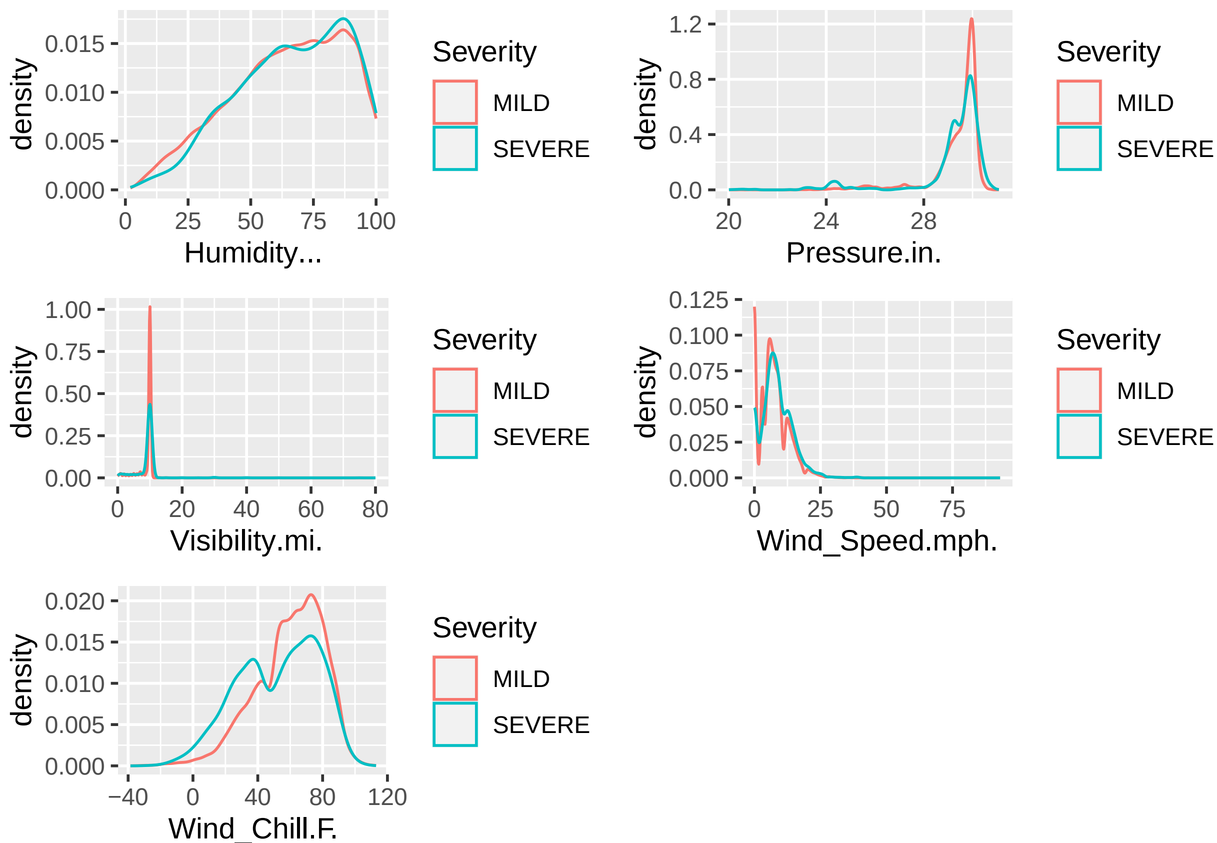
```
ggstart_lat <- ggplot(acc.train.1, aes(Start_Lat, group = Severity, color = Severity  )) + geom_density
ggstart_lng <- ggplot(acc.train.1, aes(Start_Lng, group = Severity, color = Severity  )) + geom_density
ggend_lat <- ggplot(acc.train.1, aes(End_Lat, group = Severity, color = Severity  )) + geom_density()
ggend_lng <- ggplot(acc.train.1, aes(End_Lng, group = Severity, color = Severity  )) + geom_density()
ggdistance <- ggplot(acc.train.1, aes(Distance.mi., group = Severity, color = Severity  )) + geom_densit
ggtemperature <- ggplot(acc.train.1, aes(Temperature.F., group = Severity, color = Severity  )) + geom_
gghumidity <- ggplot(acc.train.1, aes(Humidity..., group = Severity, color = Severity  )) + geom_densit
ggpressure <- ggplot(acc.train.1, aes(Pressure.in., group = Severity, color = Severity  )) + geom_densi
ggvisibility <- ggplot(acc.train.1, aes(Visibility.mi., group = Severity, color = Severity  )) + geom_de
ggwind_speed <- ggplot(acc.train.1, aes(Wind_Speed.mph., group = Severity, color = Severity  )) + geom_
ggwind_chill <- ggplot(acc.train.1, aes(Wind_Chill.F., group = Severity, color = Severity  )) + geom_de

library(gridExtra)
grid.arrange(ggstart_lat, ggstart_lng, ggend_lat)
```



```
grid.arrange(ggend_lng, ggdistance, ggtemperature)
```



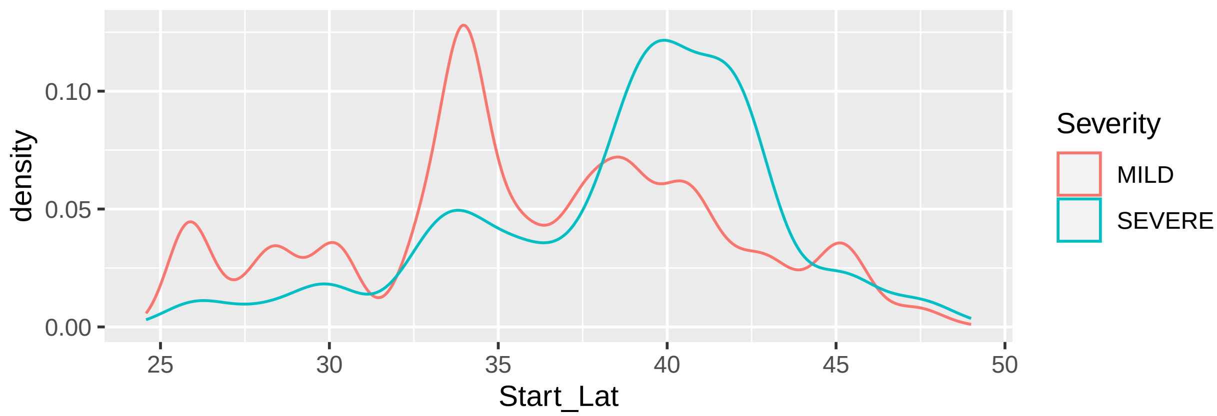

```
grid.arrange(gghumidity, ggpressure, ggvisibility, ggwind_speed, ggwind_chill)
```

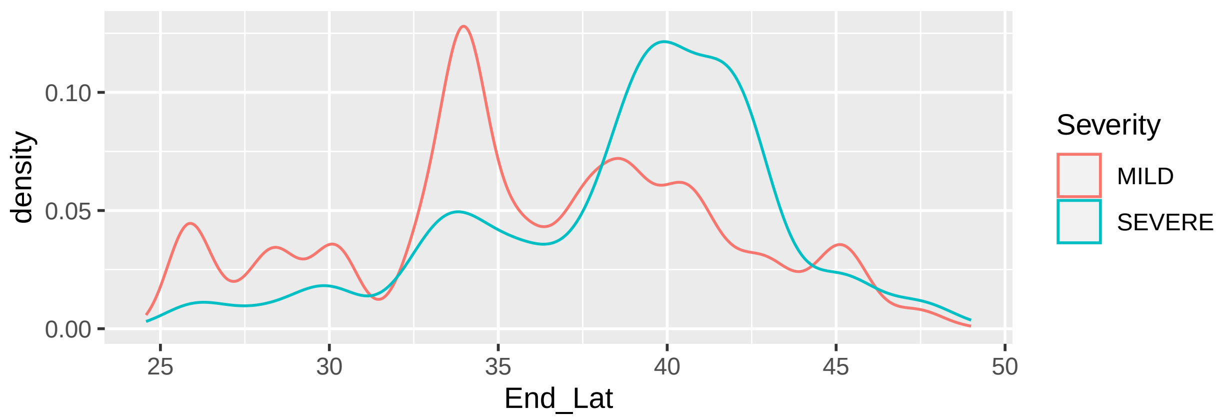


Based on the above graphs, we have that the 6 best numerical variables are Start_Lat, End_Lat, Start_Lng, End_Lng, Wind_Chill.F., Temperature.F.

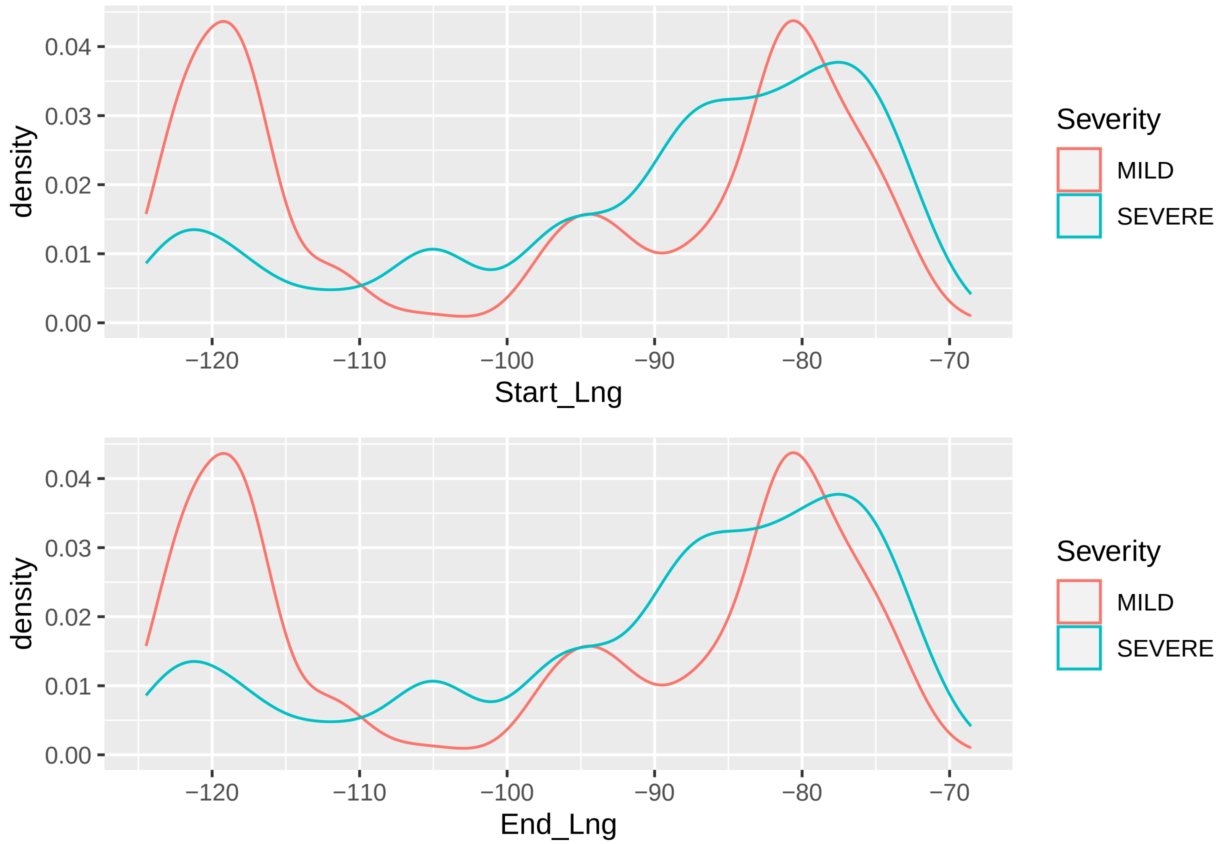
We will continue to plot these 6 numerical variables

```
grid.arrange(ggstart_lat, ggend_lat)
```

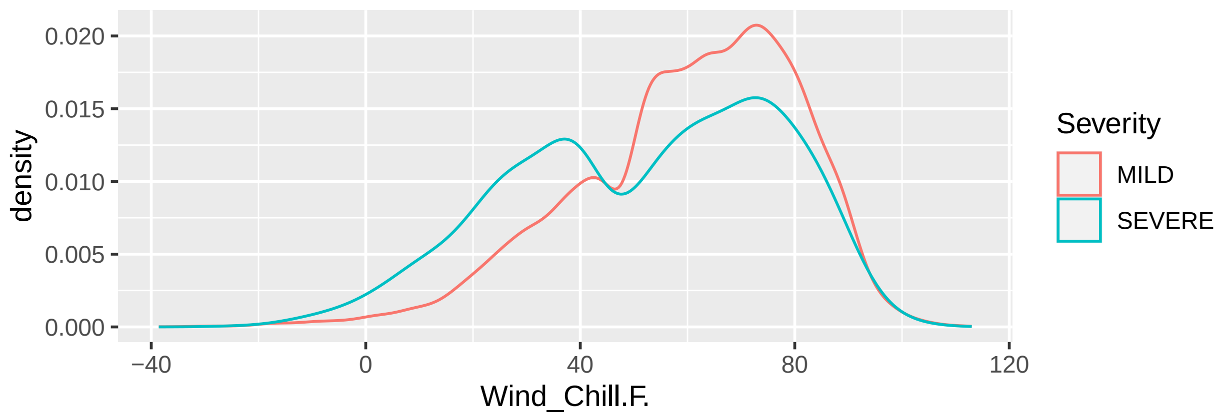


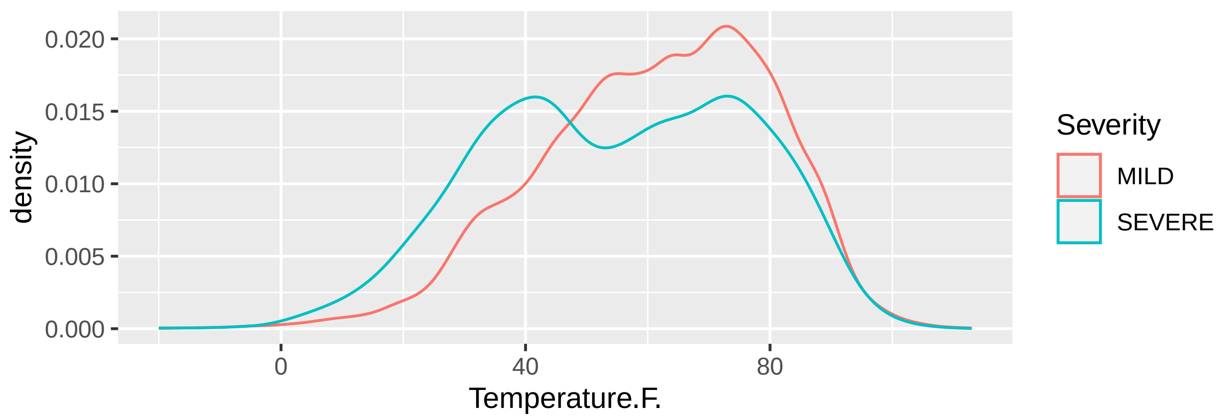



```
grid.arrange(ggstart_lng, ggend_lng)
```




```
grid.arrange(ggwind_chill, ggtemperature)
```

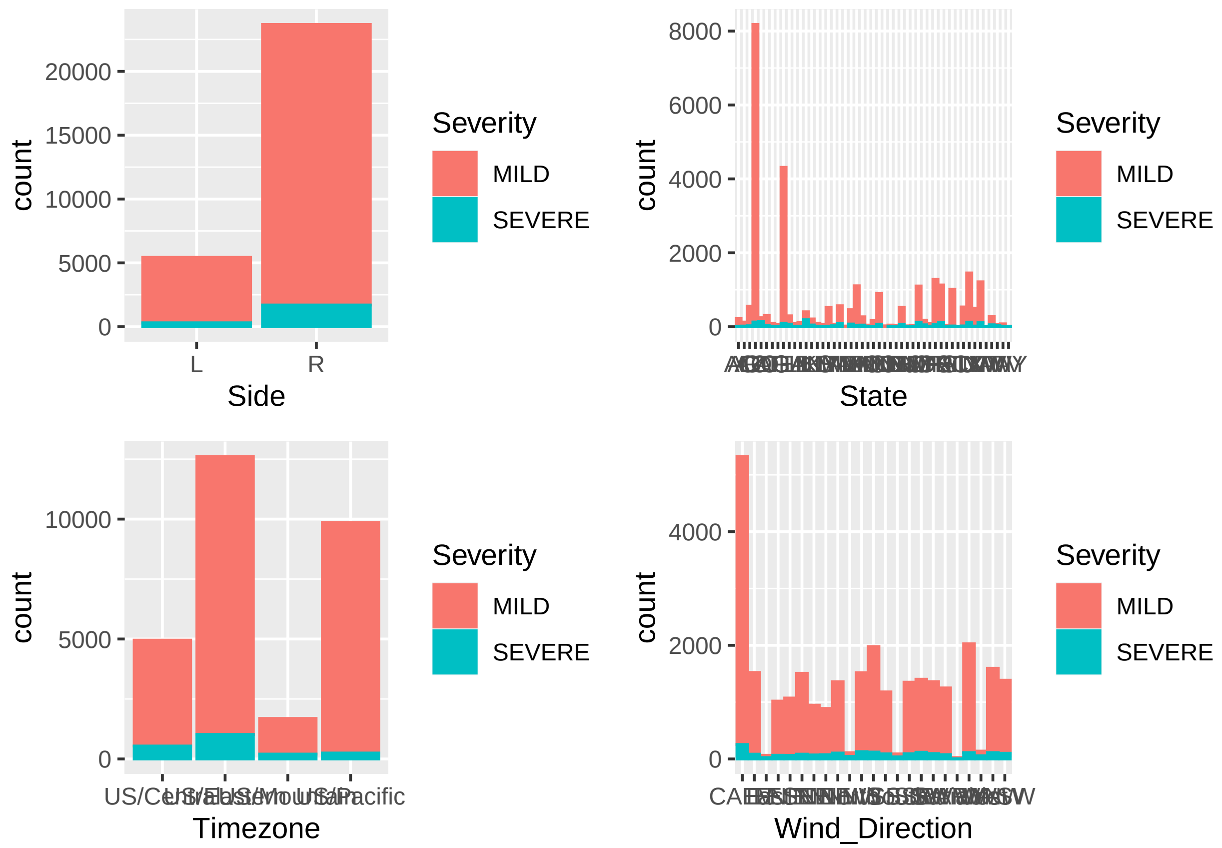




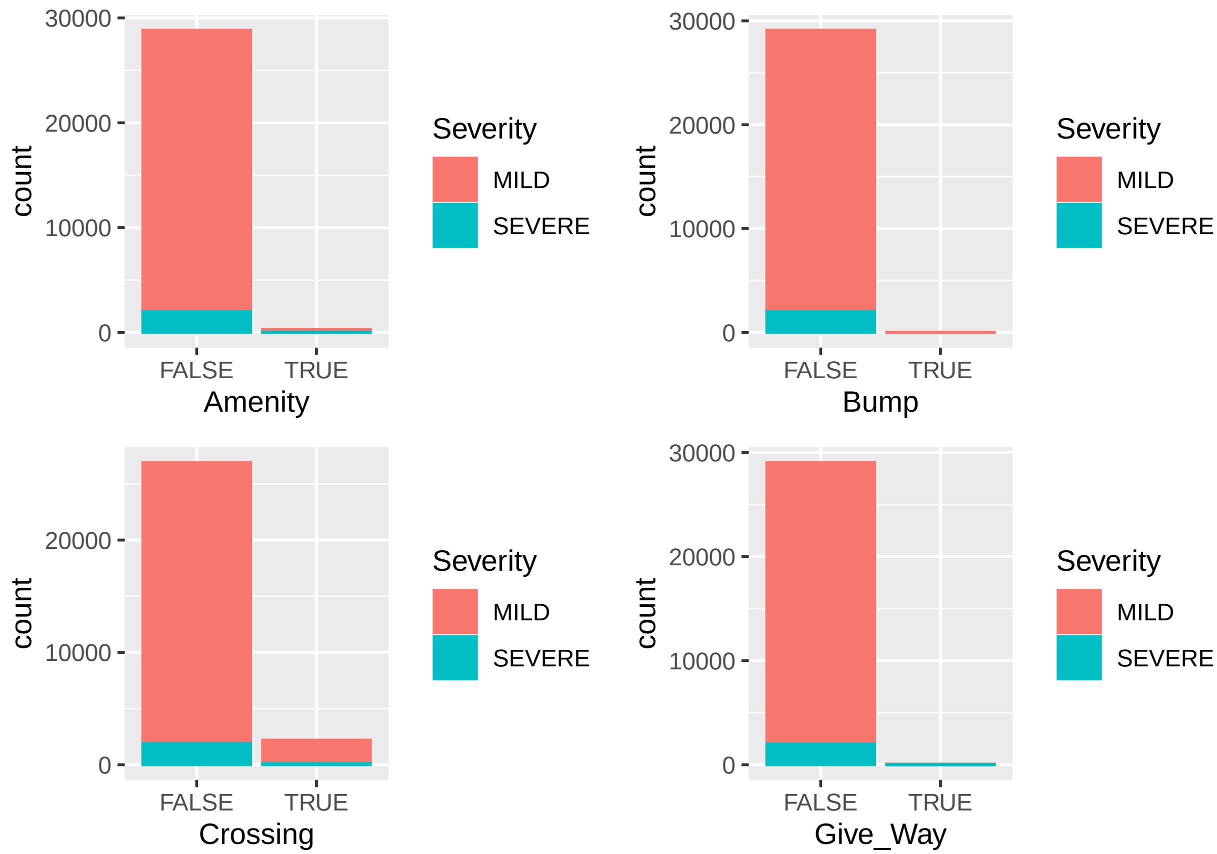
(f) Create stacked par charts for your best three categorical predictors based on your response variable.

```
library(ggplot2)
#
# Street, Side, City, Country, State, Zipcode, Country, Timezone, Airport_Code, Wind_Direction, Weather

# However, we will need to determine which predictors makes sense in the first place in the context of

ggside <- ggplot(acc.train.1, aes(Side, group = Severity, color = Severity  , fill = Severity)) + geom_
ggstate <- ggplot(acc.train.1, aes(State, group = Severity, color = Severity  , fill = Severity)) + geo
ggtimezone <- ggplot(acc.train.1, aes(Timezone, group = Severity, color = Severity  , fill = Severity))
ggwinddirection <- ggplot(acc.train.1, aes(Wind_Direction, group = Severity, color = Severity  , fill =
ggamenity <- ggplot(acc.train.1, aes(Amenity, group = Severity, color = Severity  , fill = Severity)) +
ggbump <- ggplot(acc.train.1, aes(Bump, group = Severity, color = Severity , fill = Severity )) + geom_
ggcrossing <- ggplot(acc.train.1, aes(Crossing, group = Severity, color = Severity , fill = Severity ))
gggiveway <- ggplot(acc.train.1, aes(Give_Way, group = Severity, color = Severity , fill = Severity )) +
ggjunction <- ggplot(acc.train.1, aes(Junction, group = Severity, color = Severity , fill = Severity ))
ggnoexit <- ggplot(acc.train.1, aes(No_Exit, group = Severity, color = Severity , fill = Severity )) +
ggrailway <- ggplot(acc.train.1, aes(Railway, group = Severity, color = Severity  , fill = Severity)) +
ggroundabout <- ggplot(acc.train.1, aes(Roundabout, group = Severity, color = Severity  , fill = Severi
ggstation <- ggplot(acc.train.1, aes(Station, group = Severity, color = Severity , fill = Severity )) +
ggstop <- ggplot(acc.train.1, aes(Stop, group = Severity, color = Severity , fill = Severity )) + geom_
ggcalm <- ggplot(acc.train.1, aes(Traffic_Calming, group = Severity, color = Severity  , fill = Severit
```

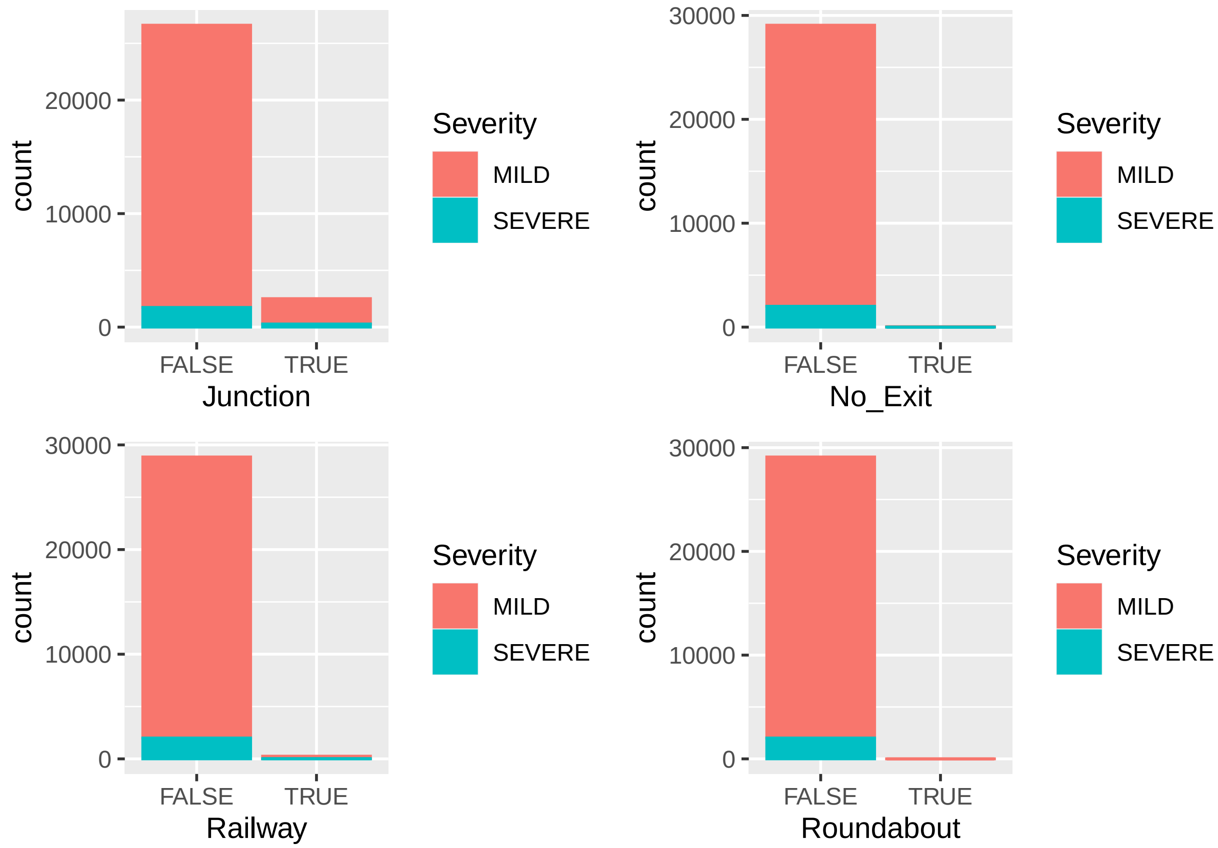
```
ggsignal <- ggplot(acc.train.1, aes(Traffic_Signal, group = Severity, color = Severity , fill = Severit
ggloop <- ggplot(acc.train.1, aes(Turning_Loop, group = Severity, color = Severity , fill = Severity ))
ggsunset<- ggplot(acc.train.1, aes(Sunrise_Sunset, group = Severity, color = Severity  , fill = Severit
ggcivil <- ggplot(acc.train.1, aes(Civil_Twilight, group = Severity, color = Severity , fill = Severity
ggnautical <- ggplot(acc.train.1, aes(Nautical_Twilight, group = Severity, color = Severity , fill = Se
ggastronomical <- ggplot(acc.train.1, aes(Astronomical_Twilight, group = Severity, color = Severity , f


library(gridExtra)
grid.arrange(ggside, ggstate, ggtimezone,ggwinddirection)
```



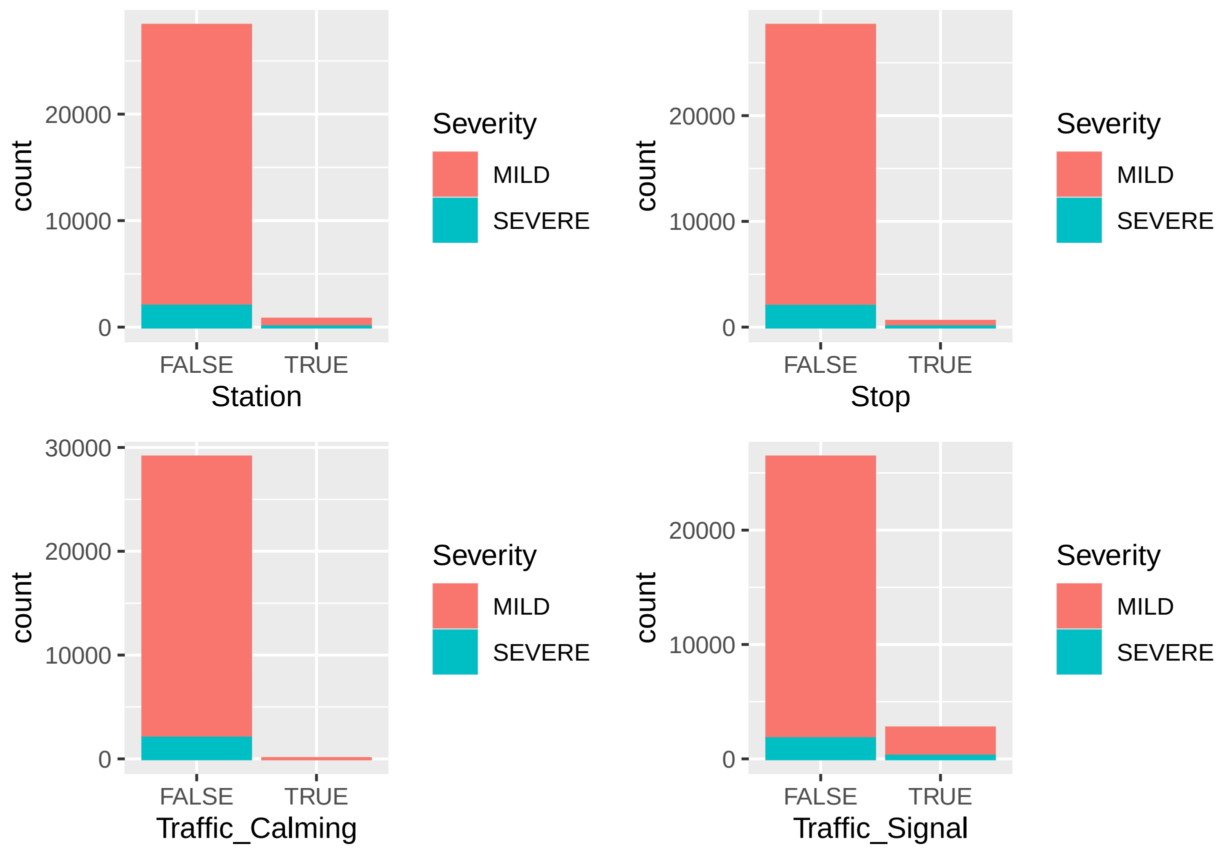

```
grid.arrange(ggamenity, ggbump,ggcrossing, gggiveway)
```



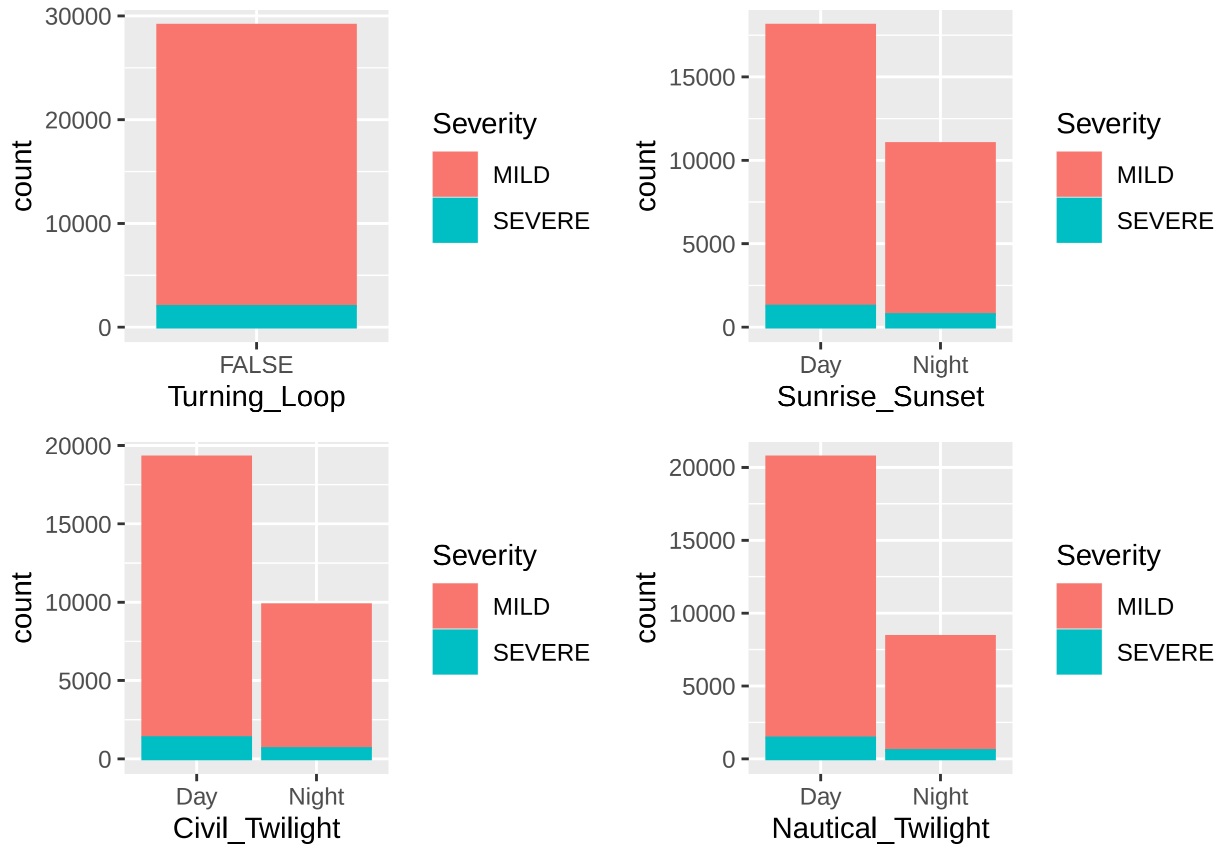

```
grid.arrange(ggjunction, ggnoexit, ggrailway, ggroundabout)
```



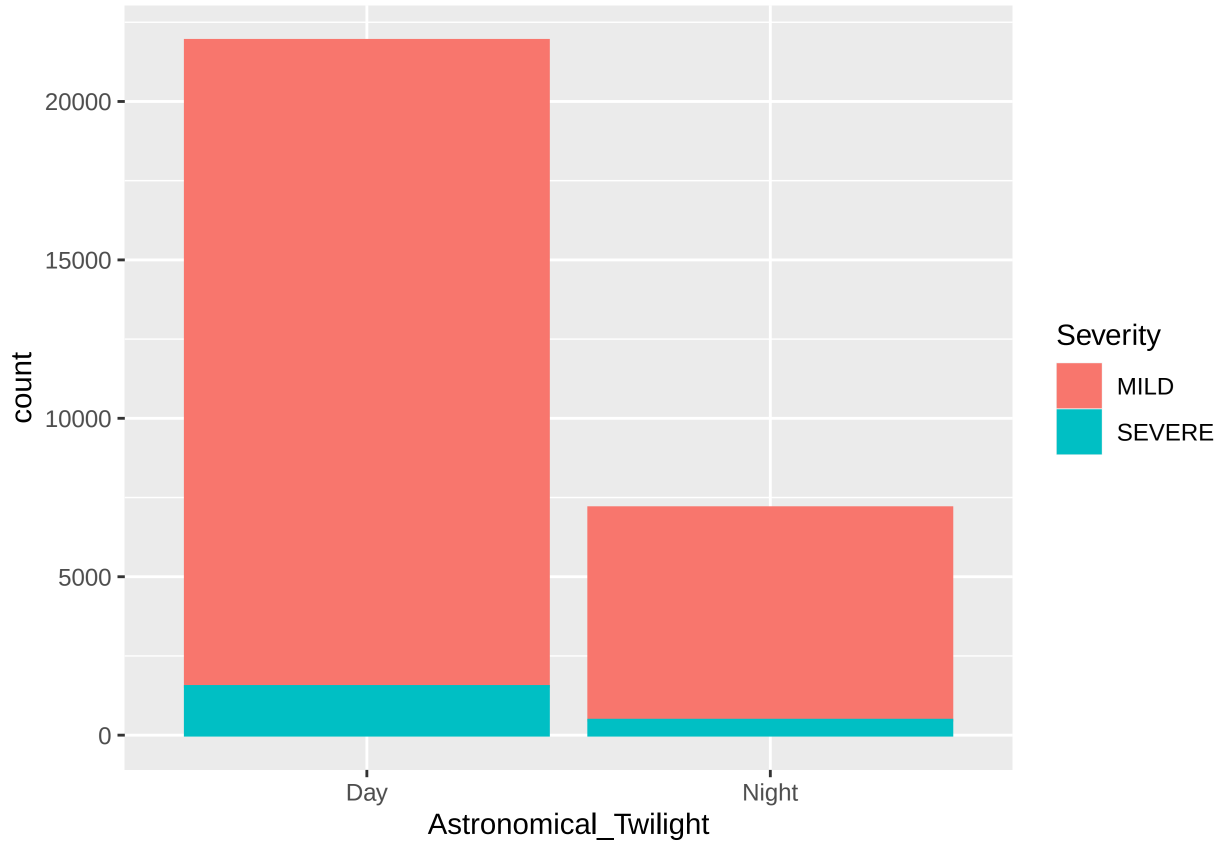

```
grid.arrange(ggstation, ggstop, ggcalm, ggsignal)
```




```
grid.arrange(ggloop, ggsunset, ggcivil, ggnautical)
```




```
grid.arrange(ggastronomical)
```



We aim to look at the proportion of the Severity (either MILD or SEVERE) between the different bars. We want to see the proportion of the MILD and SEVERE to be different between the bars and will look at these proportions to determine the best predictors.

The best categorical predictors are Timezone, Give_Way, and Traffic_Signal.

Q2

(a) Build a classifier of your choice and predict the class of the unknown Y variable "Severity" in the testing data. Create a submission file (similar to the submission file example and submit your prediction on kaggle. If you already have a group, each member must submit his/her own file.

```
numericaltest <- acc.test[,c(3,4,5,6,7,19,20,21,22,23,25)]
numericaltrain <- acc.train.1[,c(1, 4,5,6,7,8,20,21,22,23,24,26)]

# We will be imputing utilizing median values since mean can potentially draw NA's
median1 <- apply(numericaltest, 2, median, na.rm = TRUE)

dim(numericaltest)[2]
```

```
## [1] 11
```

```r
# Imputing the corresponding median values, we have that

for (i in 1:dim(numericaltest)[2]){
  for (j in 1:dim(numericaltest)[1]){
    if(is.na(numericaltest[j,i]) == TRUE){
      numericaltest[j,i] <- as.numeric(median1[i])
    }
  }
}

# pred <- acc.train$Severity
# # We will be removing the na values from both the testing and the training data sets.
#
# numericaltest.1 <- na.omit(numericaltest)
# numericaltrain.1 <- na.omit(numericaltrain)
#
pred <- numericaltrain[,1]
numericaltrain <- numericaltrain[,-1]
#
# # temp <- which(is.na(acc.train))
# # pred <- as.data.frame(pred[-temp,])
#
# length(pred)
#
# dim(numericaltest.1)
# dim(numericaltrain.1)
length(pred)
```

```
## [1] 29107
```

```r
dim(numericaltrain)
```

```
## [1] 29107    11
```

```r
sum(is.na(numericaltest))
```

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

```r
# sum(is.na(numericaltrain.1))
```

We transition into generating a classifier. This past weeks, we have showed a great emphasis on KNN models, so I will be applying what we have learned thus far.

```r
library(class)
```

```
## Warning: package 'class' was built under R version 4.1.2
```

```r
knn.model <- knn(numericaltrain, numericaltest,cl = pred, k = 1)
# knn.model
table(knn.model)
```

```
## knn.model
##   MILD SEVERE
##  13988   1012
```

```
# write.csv(knn.model,"knn.model.csv")
```

We can see that the model utilized the training and testing data to predict the severity of the accidents

(b) Report your training model (summary)

```
summary(knn.model)
```

```
##   MILD SEVERE
##  13988   1012
```

Similar to part (a), we can see the results in the table above.

(c) Report your accuracy based on your training data.

```
# With the training models, we have that

knn.train <- knn(numericaltrain, numericaltrain, cl = pred, k = 1)

# We have the confusion matrix given below
table(knn.train, pred)
```

```
##            pred
## knn.train  MILD SEVERE
##     MILD   27098      0
##     SEVERE     0   2009
```

```
# Further, we will report the misclassification rate as follows
mean(knn.train != pred)
```

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

Further, the accuracy rate is one minus the misclassification rate so we will have 1 - 0 = 1. The accuracy rate is1.

(d) Report your accuracy based on your testing (public score) on kaggle

The accuracy based on kaggle is 0.86151

(e) Report your rank on kaggle at the time the predictions were submitted based on your public score.

My prediction is 24th out of the 108 submissions that were made.

Q3

Download the birthsnewone.csv posted on bruinlearn week 6: The Y variable is the weight of the baby in grams

```
birthsnew <- read.csv("/Users/takaooba/Downloads/birthsnewone.csv")
birthsnew <- birthsnew[,-1]
```

(a)

Fit a multiple linear model using the Least Squares Approach (lm function). Report your findings.

```
set.seed(1128)
model1 <- lm(Birth.Weight..g. ~ ., data = birthsnew)
summary(model1)
```

```
##
## Call:
## lm(formula = Birth.Weight..g. ~ ., data = birthsnew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -361.06 -111.83    1.04  110.19  647.04
##
## Coefficients: (4 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  113.630103  70.652080   1.608  0.10781
## Institution.type              -5.793001   4.112440  -1.409  0.15898
## Plurality.of.birth           -45.619579   9.005894  -5.066 4.17e-07 ***
## Gender                        -9.138465   3.147681  -2.903  0.00370 **
## Race.of.child                 -0.670663   3.304553  -0.203  0.83918
## RaceOther                     16.159348  13.816763   1.170  0.24222
## RaceWhite                     20.212004   9.504359   2.127  0.03348 *
## Age.of.father                 -0.505934   0.350712  -1.443  0.14918
## Age.of.mother                  0.885874   0.441311   2.007  0.04474 *
## Education.of.father..years.    1.311397   0.816684   1.606  0.10837
## Education.of.mother..years.   -0.293253   0.871284  -0.337  0.73645
## Total.Preg                     1.608333   2.168599   0.742  0.45832
## BDead                         12.209198  12.353121   0.988  0.32301
## Terms                         -1.312101   3.625692  -0.362  0.71744
## Date.LBirth                    0.005006   0.004744   1.055  0.29132
## Month.LBirth                 -50.236160  47.676855  -1.054  0.29206
## Year.LBirth                          NA         NA      NA       NA
## LOutcome                       0.178140   1.101207   0.162  0.87149
## Weeks                         10.663338   0.773040  13.794  < 2e-16 ***
## Prenatal                       1.542984   2.316066   0.666  0.50530
## Trimester.Prenatal             5.204143   7.169221   0.726  0.46792
```

```
## Visits                        1.093058   0.462330   2.364  0.01809 *
## Birth.weight.group          454.118301   1.869896 242.857  < 2e-16 ***
## Marital                      -8.555404   4.165291  -2.054  0.04001 *
## Birth.Attendant               1.239004   2.480073   0.500  0.61738
## Numchild                            NA         NA      NA       NA
## Month.Term                    0.758097   0.837073   0.906  0.36515
## Year.Term                    -0.002471   0.003447  -0.717  0.47336
## Low.BirthNorm                20.857739   8.009332   2.604  0.00923 **
## RaceMom                             NA         NA      NA       NA
## RaceDad                      -3.924263   2.776333  -1.413  0.15756
## Mother.MinorityWhite                NA         NA      NA       NA
## Father.MinorityWhite         -6.396834   9.478744  -0.675  0.49978
## HispMomM                    -38.915071  51.140326  -0.761  0.44671
## HispMomN                    -40.024623  50.299538  -0.796  0.42622
## HispMomO                    -77.721620  61.040619  -1.273  0.20296
## HispMomP                    -35.992761  53.600230  -0.672  0.50192
## HispMomS                    -66.354014  51.505273  -1.288  0.19768
## HispMomU                     57.118474  89.186888   0.640  0.52191
## HispDadM                     -2.831209  48.063482  -0.059  0.95303
## HispDadN                     -9.641996  47.484435  -0.203  0.83910
## HispDadO                      2.828989  58.213494   0.049  0.96124
## HispDadP                    -13.091633  50.266620  -0.260  0.79453
## HispDadS                     24.339338  48.388548   0.503  0.61498
## HispDadU                    -79.559371  82.246593  -0.967  0.33341
## AveCigs                       1.695141   0.896925   1.890  0.05880 .
## SmokerNo                     39.508524  10.024921   3.941 8.18e-05 ***
## AveDrink                      4.645455  12.954389   0.359  0.71990
## Wt.Gain                       0.561398   0.119590   4.694 2.72e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 137.7 on 7816 degrees of freedom
## Multiple R-squared:  0.9491, Adjusted R-squared:  0.9488
## F-statistic:  3314 on 44 and 7816 DF,  p-value: < 2.2e-16
```

Based on the linear model, we have that the significant predictors are Plurality.of.birth, gender, RaceWhite, Age.of.mother, Weeks, Visits, Birth.weight.group, Marital, Low.BirthNorm, SmokerNo, and Wt.Gain.

(b) Use Ridge Regression Approach to predict the weight of the baby in grams. Interpret the resulting model.

```r
# install.packages("glmnet")
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.1.2
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-4
```

```
head(birthsnew)
```

```
##   Institution.type Plurality.of.birth Gender Race.of.child  Race Age.of.father
## 1                1                  1      2             1 White            50
## 2                1                  1      2             1 White            19
## 3                1                  1      2             1 White            37
## 4                1                  1      2             2 Black            39
## 5                1                  1      1             2 Black            20
## 6                1                  1      2             1 White            30
##   Age.of.mother Education.of.father..years. Education.of.mother..years.
## 1            24                          12                          15
## 2            18                           9                           9
## 3            35                          17                          17
## 4            31                          11                          16
## 5            19                          11                          12
## 6            27                          16                          16
##   Total.Preg BDead Terms Date.LBirth Month.LBirth Year.LBirth LOutcome Weeks
## 1          2     0     0       32004            3        2004        1    38
## 2          1     0     0           0            0           0        9    35
## 3          2     0     0      112003           11        2003        1    38
## 4          1     0     0           0            0           0        9    38
## 5          1     0     0           0            0           0        9    36
## 6          1     0     0           0            0           0        9    40
##   Prenatal Trimester.Prenatal Visits Birth.weight.group Marital Birth.Attendant
## 1        3                  1     10                  5       2               1
## 2        3                  1      9                  6       2               1
## 3        1                  1     20                  5       1               1
## 4        6                  2     12                  5       2               1
## 5        4                  2     10                  6       2               1
## 6        1                  1     20                  6       1               1
##   Numchild Month.Term Year.Term Low.Birth RaceMom RaceDad Mother.Minority
## 1        1          0         0      Norm       1       2           White
## 2        0          0         0      Norm       1       1           White
## 3        1          0         0      Norm       1       1           White
## 4        0          0         0      Norm       2       2        Nonwhite
## 5        0          0         0      Norm       2       1        Nonwhite
## 6        0          0         0      Norm       1       1           White
##   Father.Minority HispMom HispDad AveCigs Smoker AveDrink Wt.Gain
## 1        Nonwhite       N       N       0     No        0      50
## 2           White       N       N      23   Cigs        0      35
## 3           White       N       N       0     No        0      24
## 4        Nonwhite       N       N       0     No        0      30
## 5           White       N       M       0     No        0      10
## 6           White       N       N       0     No        0      37
##   Birth.Weight..g.
## 1         2865.875
## 2         3121.250
## 3         2667.250
## 4         2979.375
## 5         3036.125
## 6         3092.875
```

```r
x = model.matrix(Birth.Weight..g. ~ ., data = birthsnew)
y = birthsnew$Birth.Weight..g.

# As given in the problem statement, we have the following
i = seq(10, -2, length = 100)
lambda.v = 10^i

model.ridge <- glmnet(x, y, alpha = 0, lambda = lambda.v)
summary(model.ridge)
```

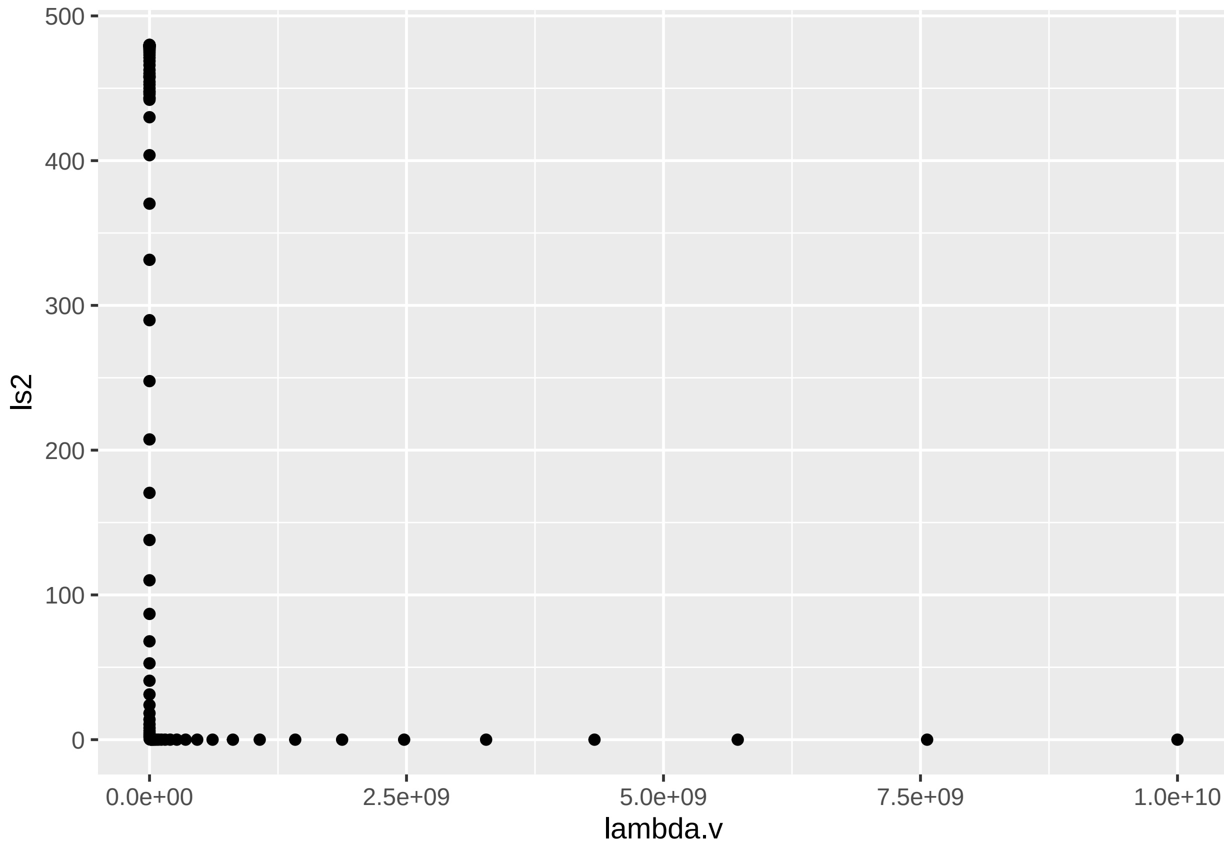
```
##           Length Class     Mode
## a0         100   -none-    numeric
## beta      4900   dgCMatrix S4
## df         100   -none-    numeric
## dim          2   -none-    numeric
## lambda     100   -none-    numeric
## dev.ratio  100   -none-    numeric
## nulldev      1   -none-    numeric
## npasses      1   -none-    numeric
## jerr         1   -none-    numeric
## offset       1   -none-    logical
## call         5   -none-    call
## nobs         1   -none-    numeric
```

```r
# Get the coef of the model
coeffs <- coef(model.ridge)
dim(coeffs)
```

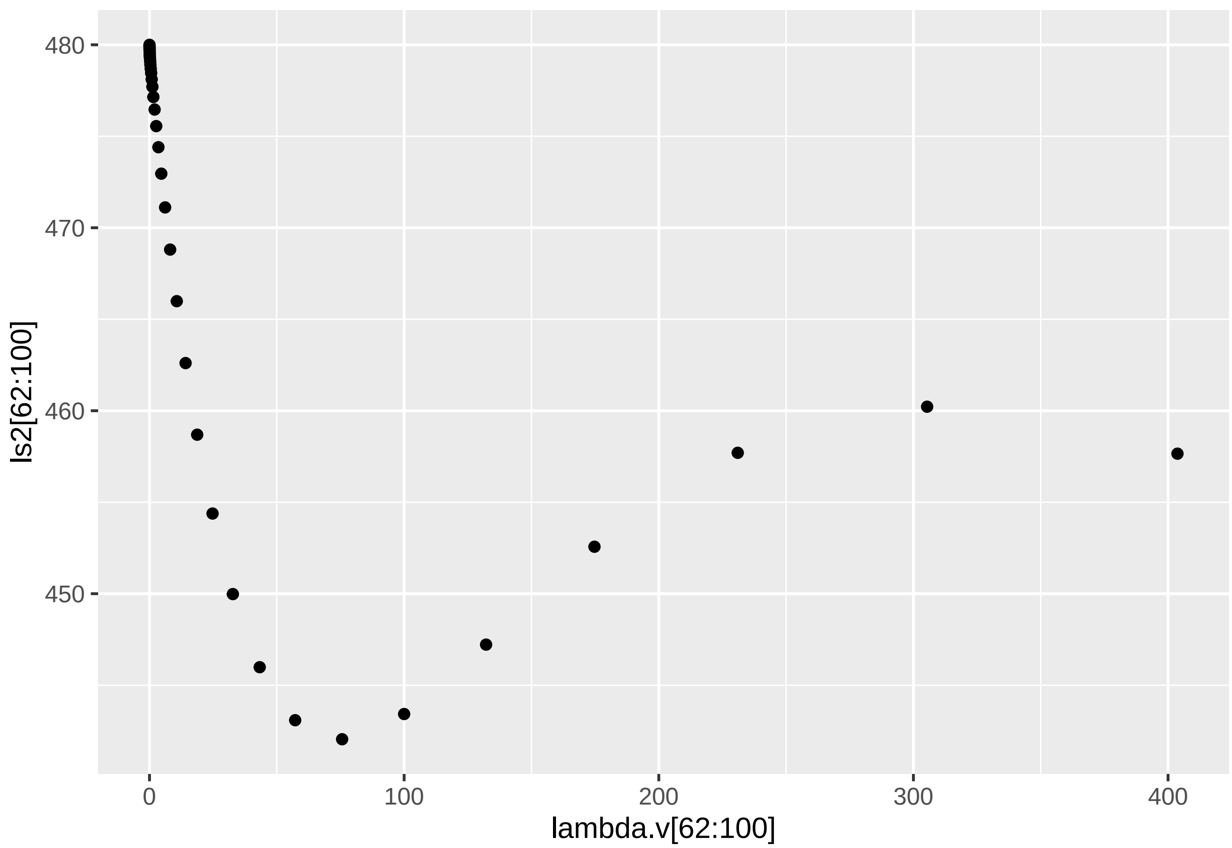
```
## [1]  50 100
```

```r
my.l2 <- function(betas)
{
  sqrt(sum(betas^2))
}
ls2 <- c()
for (i in 1:100){ls2 = c(ls2, my.l2(coeffs[-c(1,2),i]))}



qplot(lambda.v, ls2)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```




```
# We look at tthe graph more closely
qplot(lambda.v[62:100], ls2[62:100])
```




```
set.seed(1128)
cv.output = cv.glmnet(x,y, alpha = 0)
bestreg.cvL = cv.output$lambda.min
bestreg.cvL
```

```
## [1] 59.2054
```

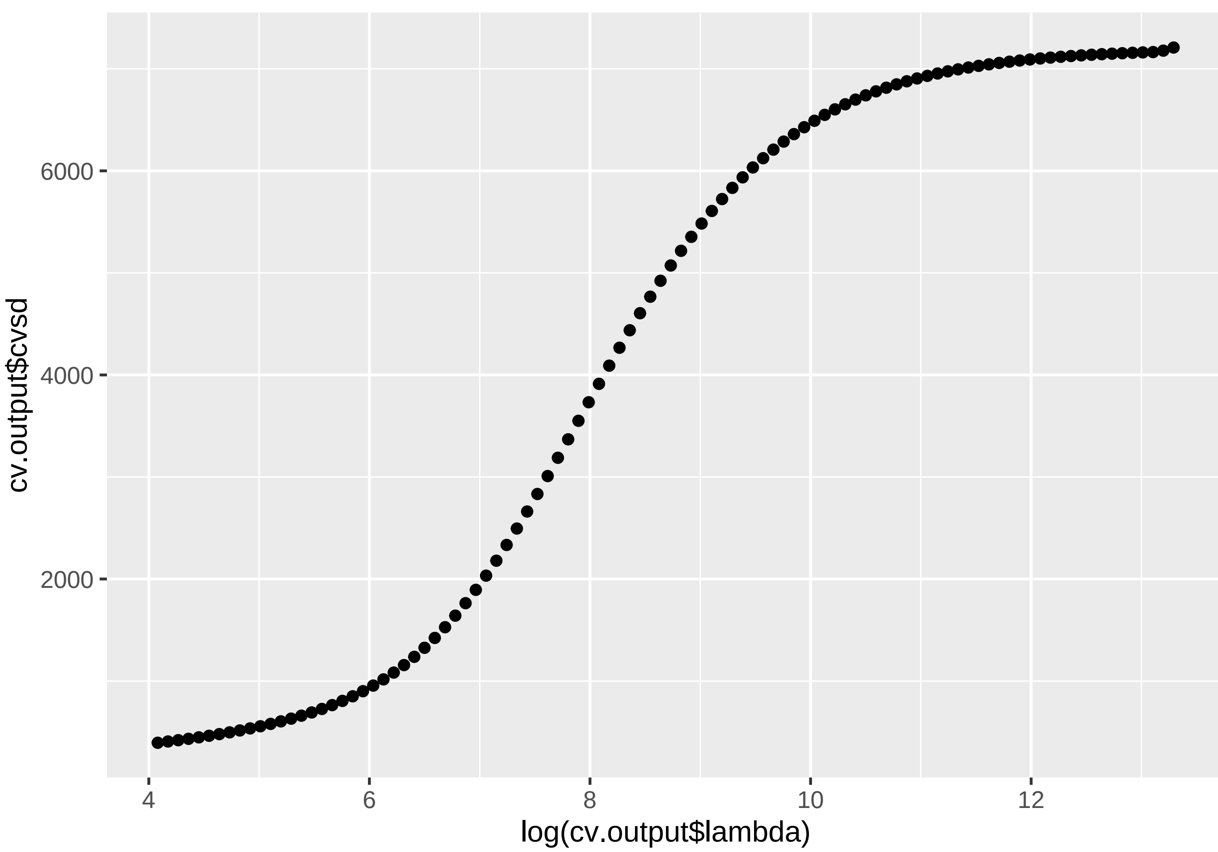
```
bestcoeff = predict(model.ridge, s= bestreg.cvL, type = "coefficients")
sqrt(sum(bestcoeff^2))
```

```
## [1] 443.3752
```

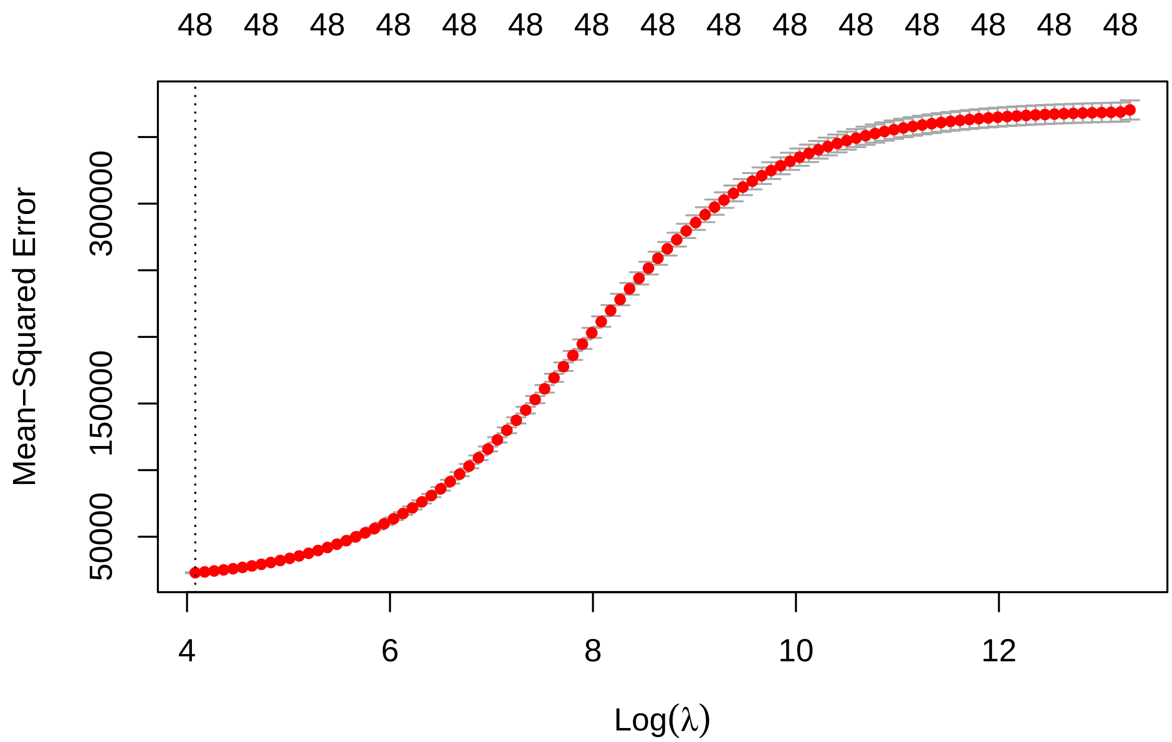
Above is the predicted weight of the baby in grams when utilizing the Ridge Regression pproach. It is the quantity of how the coefficients change as lambda values grows.

(c) Make a plot that shows how the ratio of the size of the coefficients for Ridge Regression to the size of the coefficients for LS Change as lambda gets bigger. Your x-axis should have the values of lambda (from 10(-2) to 10(10). The y-axis should have the ratio of the L2 norm of the Ridge Regression coefficients divided by the L2 Norm of the Least Squares coefficients. (Hint: you'll need to do a LS Regression with the variables standardized. Do not drop any terms from the model.)

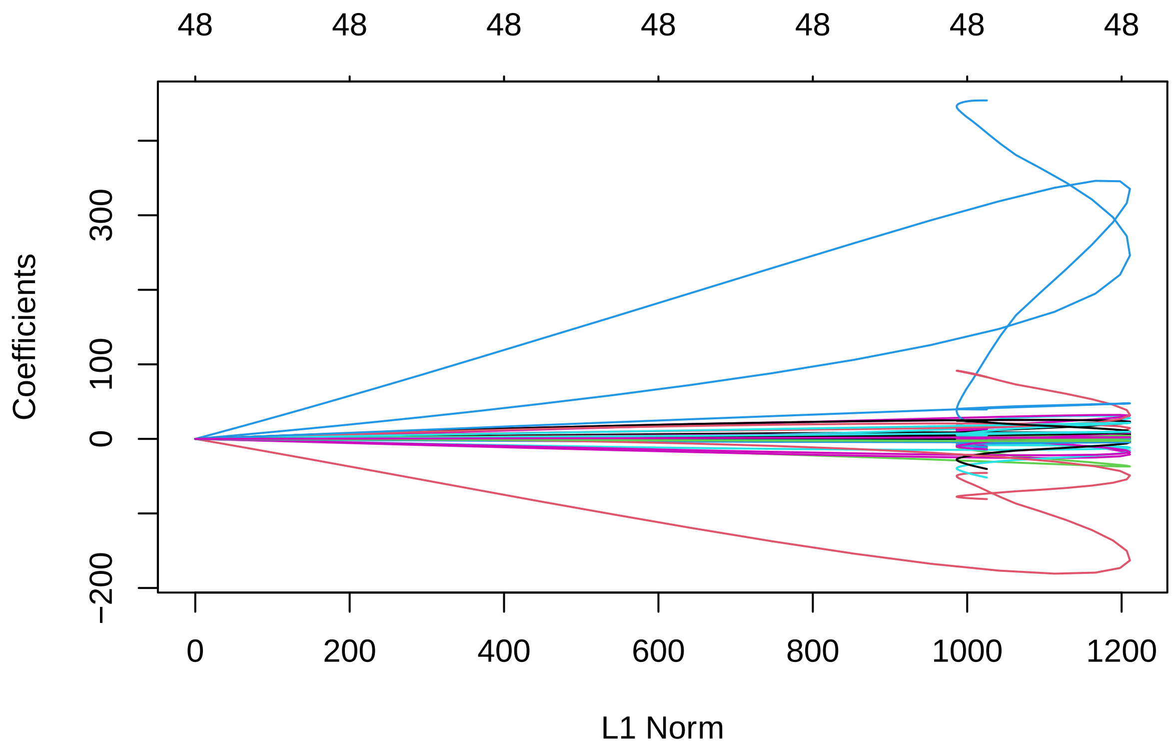
```
qplot(log(cv.output$lambda), cv.output$cvsd)
```




```
plot(cv.output)
```




```
plot(model.ridge)
```



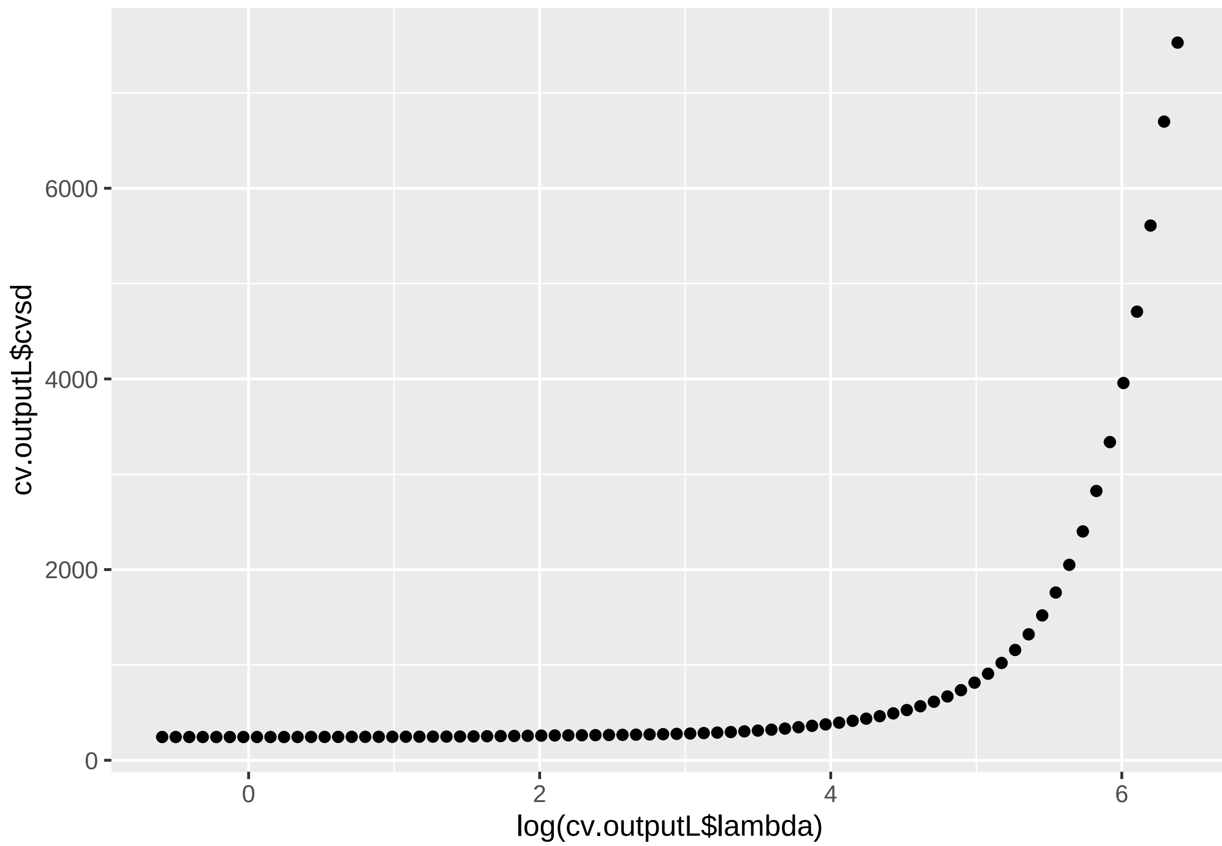
We can see that on the very top, we have that the value (48) does not decrease as we move to the right of the x axis, thus we will look at an alternative model. At the next part, we will be looking at the Lasso method

(d) Use Lasso Regression Approach to predict the weight of the baby in grams and interpret

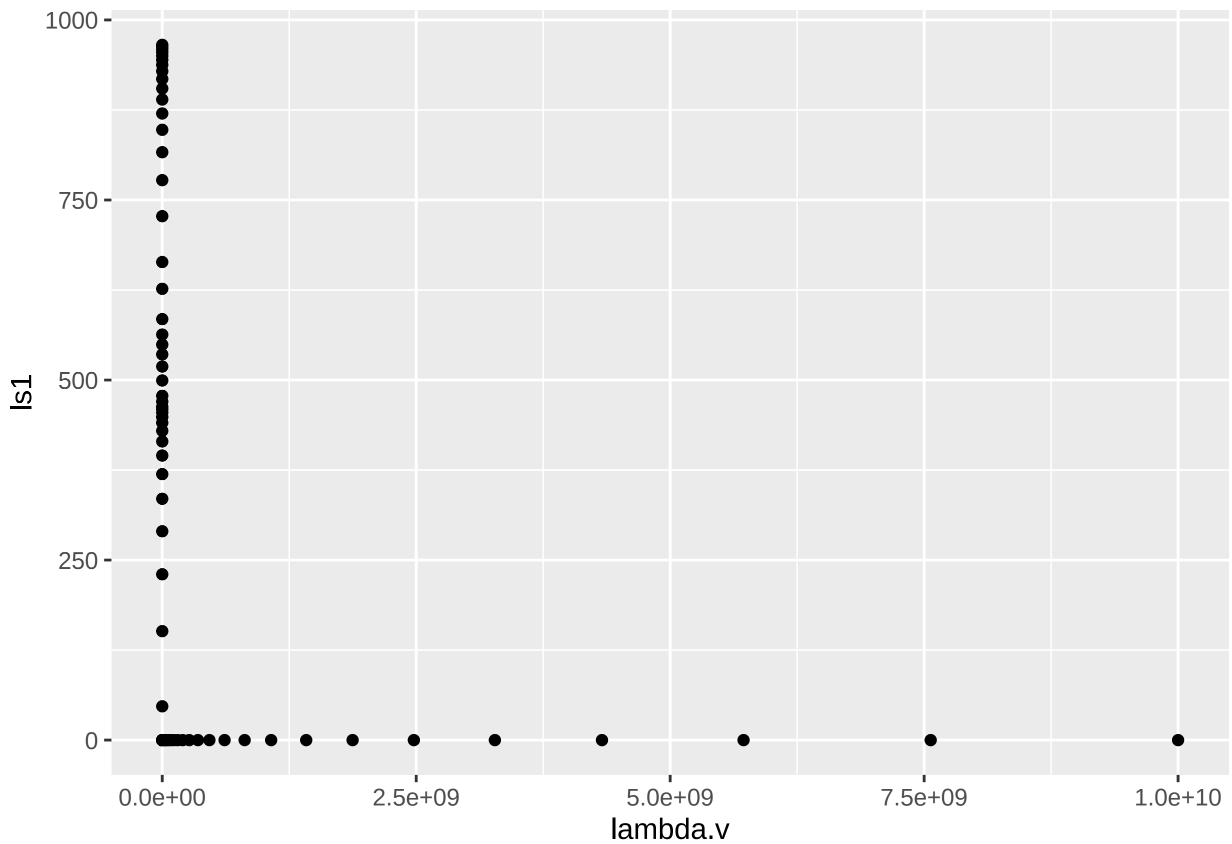
```
model.lasso <- glmnet(x,y,alpha = 1, lambda = lambda.v)

summary(model.lasso)
```

```
##           Length Class     Mode
## a0         100   -none-    numeric
## beta      4900   dgCMatrix S4
## df         100   -none-    numeric
## dim          2   -none-    numeric
## lambda     100   -none-    numeric
## dev.ratio  100   -none-    numeric
## nulldev      1   -none-    numeric
## npasses      1   -none-    numeric
## jerr         1   -none-    numeric
## offset       1   -none-    logical
## call         5   -none-    call
## nobs         1   -none-    numeric
```

```
coeffsL <- coef(model.lasso)
set.seed(1128)

cv.outputL=cv.glmnet(x,y,alpha=1)
qplot(log(cv.outputL$lambda),cv.outputL$cvsd)
```

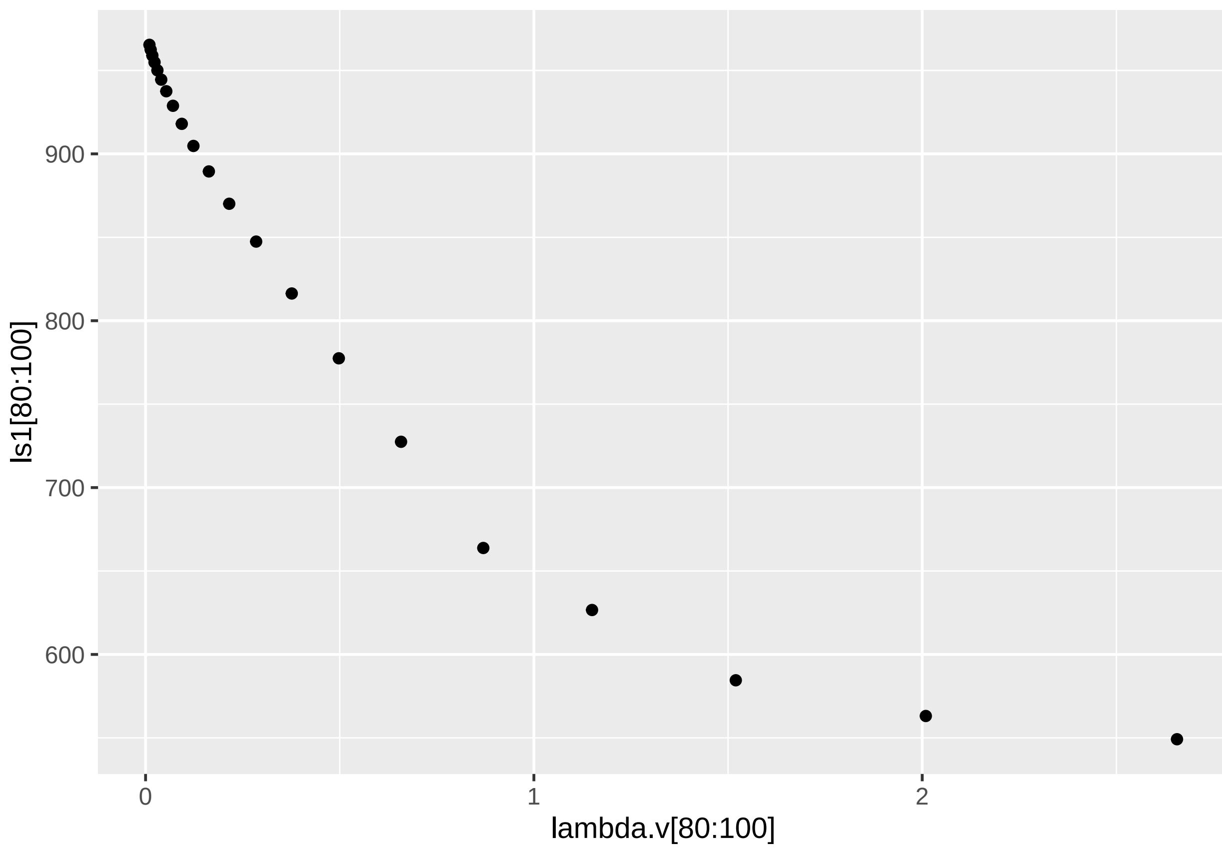



```
my.l1=function(betas){#calculate l1 norm
 sum(abs(betas))}
 ls1=c()
 for (i in 1:100){ ls1=c(ls1, my.l1(coeffsL[-c(1,2),i]))}

qplot(lambda.v,ls1)
```




```
#Zooming in we have that
qplot(lambda.v[80:100],ls1[80:100], type = "b")
```

```
## Warning in geom_point(type = "b"): Ignoring unknown parameters: `type`
```




```
cv.output = cv.glmnet(x,y, alpha = 1)
bestlamb.cvL=cv.outputL$lambda.min
bestlamb.cvL
```

```
## [1] 1.162226
```

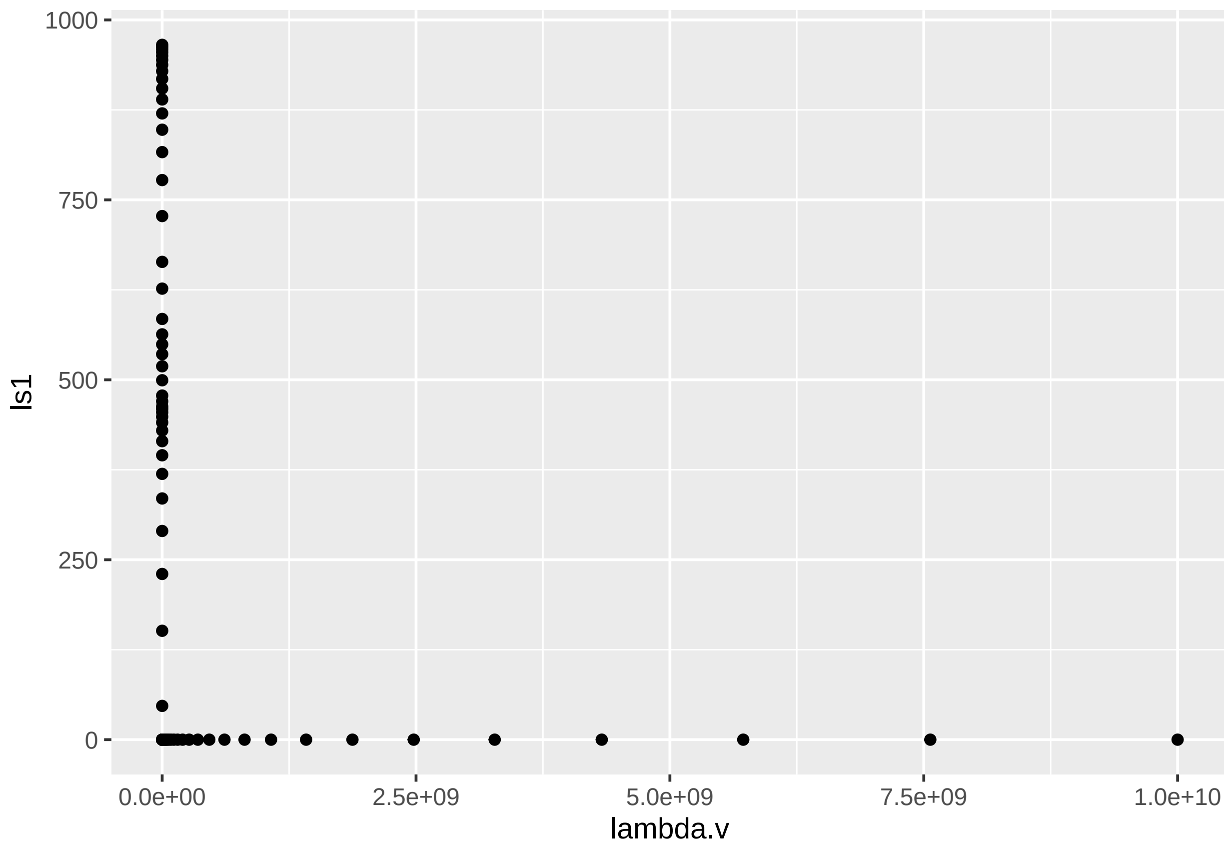
```
bestcoeffL = predict(model.lasso, s= bestlamb.cvL, type = "coefficients")
sum(abs(bestcoeffL))
```

```
## [1] 722.1132
```

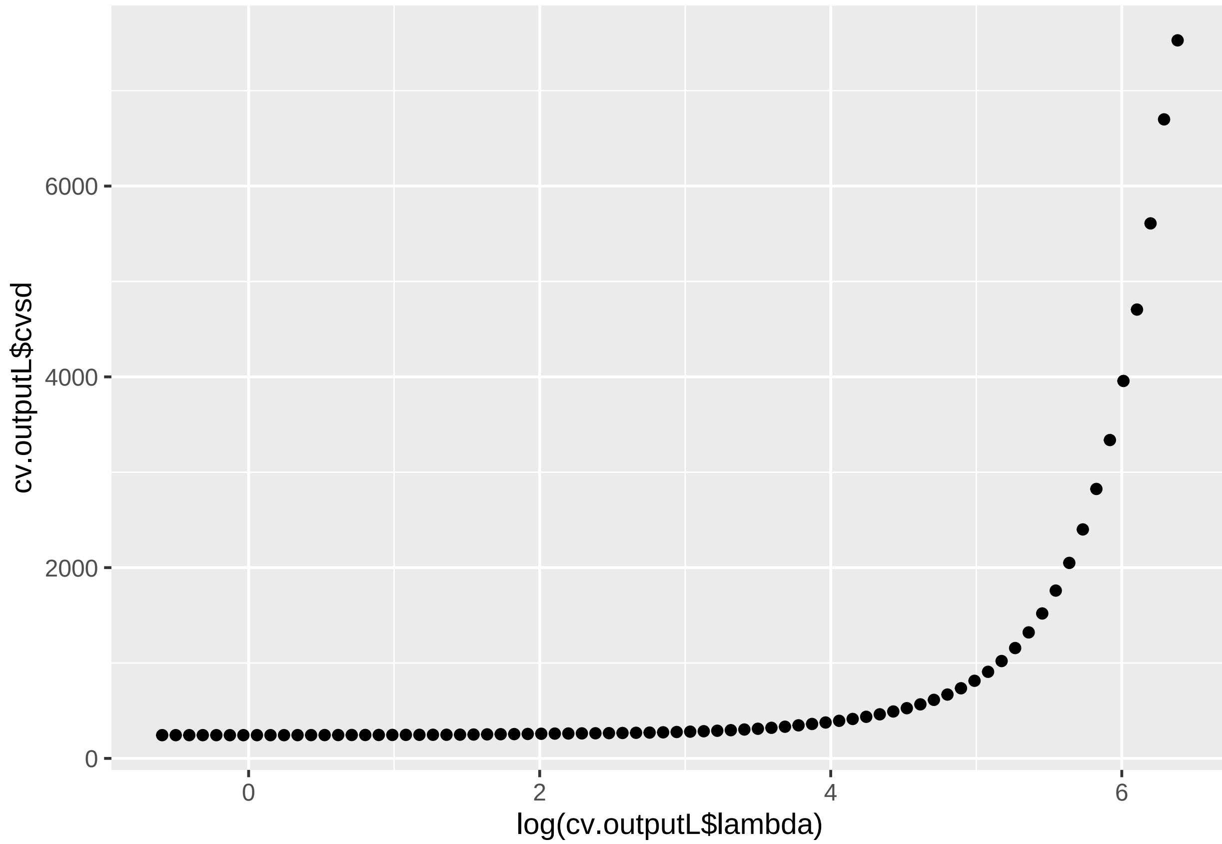
Above is the predicted weight of the baby in grams when utilizing the Lasso Regression approach. It is the quantity of how the coefficients change as lambda values grows.

(e) Repeat (c) Using Lasso Regression.

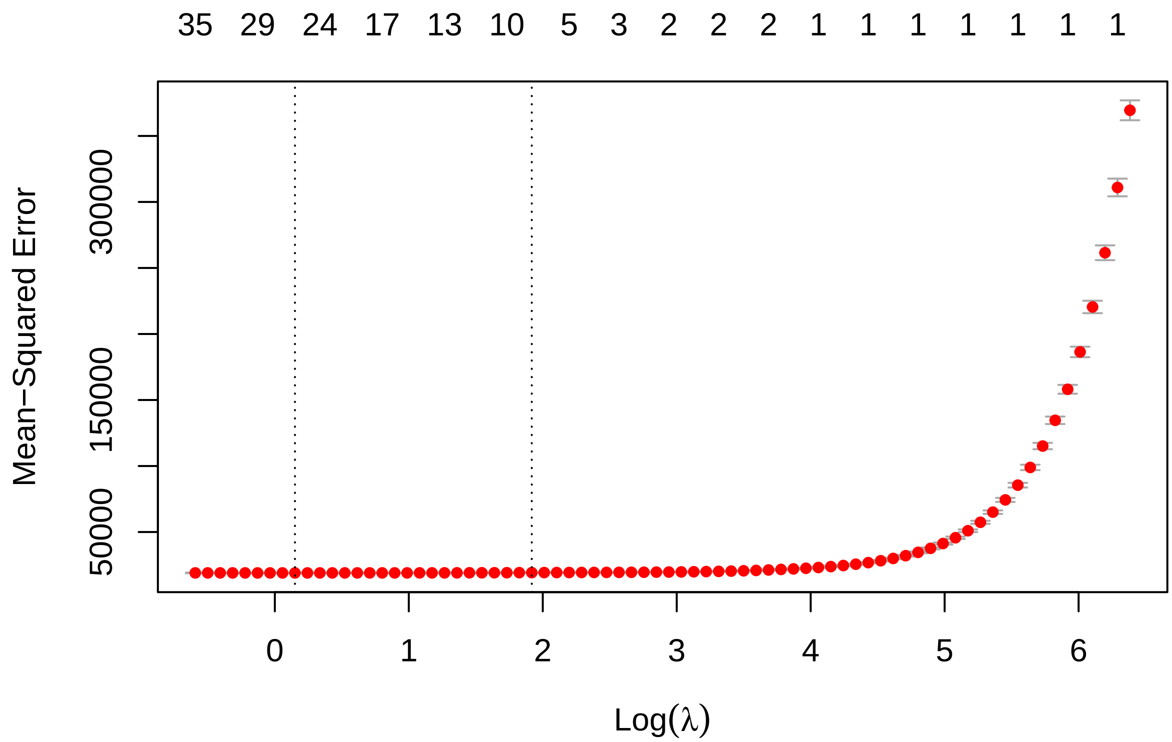
```
qplot(lambda.v, ls1)
```




```
qplot(log(cv.outputL$lambda), cv.outputL$cvsd)
```




```
plot(cv.outputL)
```



Write a short paragraph compar

Note the difference between the above graph and the graph constructed in part (c). We can see that at the very top of the graph or the top label, we note that the values are getting gradually smaller as we move to the right of the x axis. This is a good thing and was not seen in the graph constructed in part (c). We can thus conclude that the lasso regression model is a better approach then the Ridge Regression Approach.