EDAFinalProject

Kyu Min Shim

2023-04-20

Total word count: 999

Task 1

```
data = read.csv("UK_Accident.csv")
summary(data)
```

```
##        X          Accident_Index     Location_Easting_OSGR
##  Min.   :     0   Length:1504150     Min.   : 64950
##  1st Qu.:125345   Class :character   1st Qu.:375060
##  Median :250691   Mode  :character   Median :439960
##  Mean   :253043                      Mean   :439621
##  3rd Qu.:376037                      3rd Qu.:523060
##  Max.   :570010                      Max.   :655370
##                                      NA's   :101
##  Location_Northing_OSGR   Longitude          Latitude      Police_Force
##  Min.   :      0        Min.   :-7.5162   Min.   : 0.00   Min.   : 1.00
##  1st Qu.: 178260        1st Qu.:-2.3739   1st Qu.:51.49   1st Qu.: 6.00
##  Median : 268800        Median :-1.4037   Median :52.31   Median :30.00
##  Mean   : 300138        Mean   :-1.4366   Mean   :52.59   Mean   :30.21
##  3rd Qu.: 398150        3rd Qu.:-0.2215   3rd Qu.:53.48   3rd Qu.:45.00
##  Max.   :1208800        Max.   : 1.7594   Max.   :60.76   Max.   :98.00
##                         NA's   :101
##  Accident_Severity Number_of_Vehicles Number_of_Casualties     Date
##  Min.   :1.000     Min.   : 1.000     Min.   : 1.000       Length:1504150
##  1st Qu.:3.000     1st Qu.: 1.000     1st Qu.: 1.000       Class :character
##  Median :3.000     Median : 2.000     Median : 1.000       Mode  :character
##  Mean   :2.838     Mean   : 1.832     Mean   : 1.351
##  3rd Qu.:3.000     3rd Qu.: 2.000     3rd Qu.: 1.000
##  Max.   :3.000     Max.   :67.000     Max.   :93.000
##
##   Day_of_Week        Time           Local_Authority_.District.
##  Min.   :1.000   Length:1504150     Min.   :  1.0
##  1st Qu.:2.000   Class :character   1st Qu.:110.0
##  Median :4.000   Mode  :character   Median :322.0
##  Mean   :4.119                      Mean   :347.6
##  3rd Qu.:6.000                      3rd Qu.:518.0
##  Max.   :7.000                      Max.   :941.0
##
```

```
##  Local_Authority_.Highway. X1st_Road_Class X1st_Road_Number  Road_Type
##  Length:1504150            Min.   :1.000   Min.   :  -1     Length:1504150
##  Class :character          1st Qu.:3.000   1st Qu.:   0     Class :character
##  Mode  :character          Median :4.000   Median : 129     Mode  :character
##                            Mean   :4.088   Mean   :1010
##                            3rd Qu.:6.000   3rd Qu.: 725
##                            Max.   :6.000   Max.   :9999
##
##   Speed_limit    Junction_Control   X2nd_Road_Class  X2nd_Road_Number
##  Min.   :10.00   Length:1504150     Min.   :-1.000   Min.   :  -1.0
##  1st Qu.:30.00   Class :character   1st Qu.:-1.000   1st Qu.:   0.0
##  Median :30.00   Mode  :character   Median : 3.000   Median :   0.0
##  Mean   :39.01                      Mean   : 2.675   Mean   : 381.6
##  3rd Qu.:50.00                      3rd Qu.: 6.000   3rd Qu.:   0.0
##  Max.   :70.00                      Max.   : 6.000   Max.   :9999.0
##
##  Pedestrian_Crossing.Human_Control Pedestrian_Crossing.Physical_Facilities
##  Length:1504150                    Length:1504150
##  Class :character                  Class :character
##  Mode  :character                  Mode  :character
##
##
##
##
##  Light_Conditions   Weather_Conditions Road_Surface_Conditions
##  Length:1504150     Length:1504150     Length:1504150
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Special_Conditions_at_Site Carriageway_Hazards Urban_or_Rural_Area
##  Length:1504150             Length:1504150      Min.   :1.000
##  Class :character           Class :character    1st Qu.:1.000
##  Mode  :character           Mode  :character    Median :1.000
##                                                 Mean   :1.354
##                                                 3rd Qu.:2.000
##                                                 Max.   :3.000
##
##  Did_Police_Officer_Attend_Scene_of_Accident LSOA_of_Accident_Location
##  Length:1504150                              Length:1504150
##  Class :character                            Class :character
##  Mode  :character                            Mode  :character
##
##
##
##
##       Year
##  Min.   :2005
##  1st Qu.:2006
##  Median :2010
##  Mean   :2009
##  3rd Qu.:2012
```

```
##  Max.   :2014
##
```

The dataset I chose contains 33 features of 1.5 million road accidents in the UK between the years 2005 and 2014. One topic I want to consider is the relationship between accident location (given by longitude and lattitude) and the number of casualties, as well as the relationship between accident location and the number of vehicles involved. This would provide us locations in UK where accidents are more likely to occur, possibly due to the local weather/ground/social conditions, and suggest additions of more road safety measures in the area.

Another topic I want to consider is predicting accident severity based on various road conditions such as the speed-limit, road type, light conditions, weather conditions and road surface conditions. This can provide the knowledge of when drivers should take extra caution to avoid highly severe accidents. Since light conditions and weather conditions are string variables given in sentences, I would perform data mining on these variables to identify and categorize each entry by their key words.

Task 2

This section shows the code used to clean the data in preparation for other mandatory tasks. I am only interest in the most recent year data, since road conditions may have changed over time which may impact the relationship between variables. After filtering for the most recent year data (2014), I will choose the 5 variables that I believe are the most important in explaining a target variable of Accident_Severity: Number_of_Vehicles, Number_of_Casualties, Speed_Limit, Junction_Control, and Road_Type. Then, I will handle missing data by getting rid of accidents that do not have all 6 variables (including Accident_Severity) and factorizing the categorical variables. Also, column names will be shortened for better visualization in ggpairs plot.

```
myvars = c("Accident_Severity","Number_of_Vehicles","Number_of_Casualties",
           "Speed_limit","Junction_Control","Road_Type")
subdata = data[data$Year == 2014, ][myvars]
subdata = na.omit(subdata)
subdata$Junction_Control = as.factor(subdata$Junction_Control)
subdata$Road_Type = as.factor(subdata$Road_Type)
subdata$Accident_Severity = factor(subdata$Accident_Severity, levels=c(1,2,3))
colnames(subdata) = c("Severity","Cars","Casualties","Speed","Control","Road")
```

Task 3

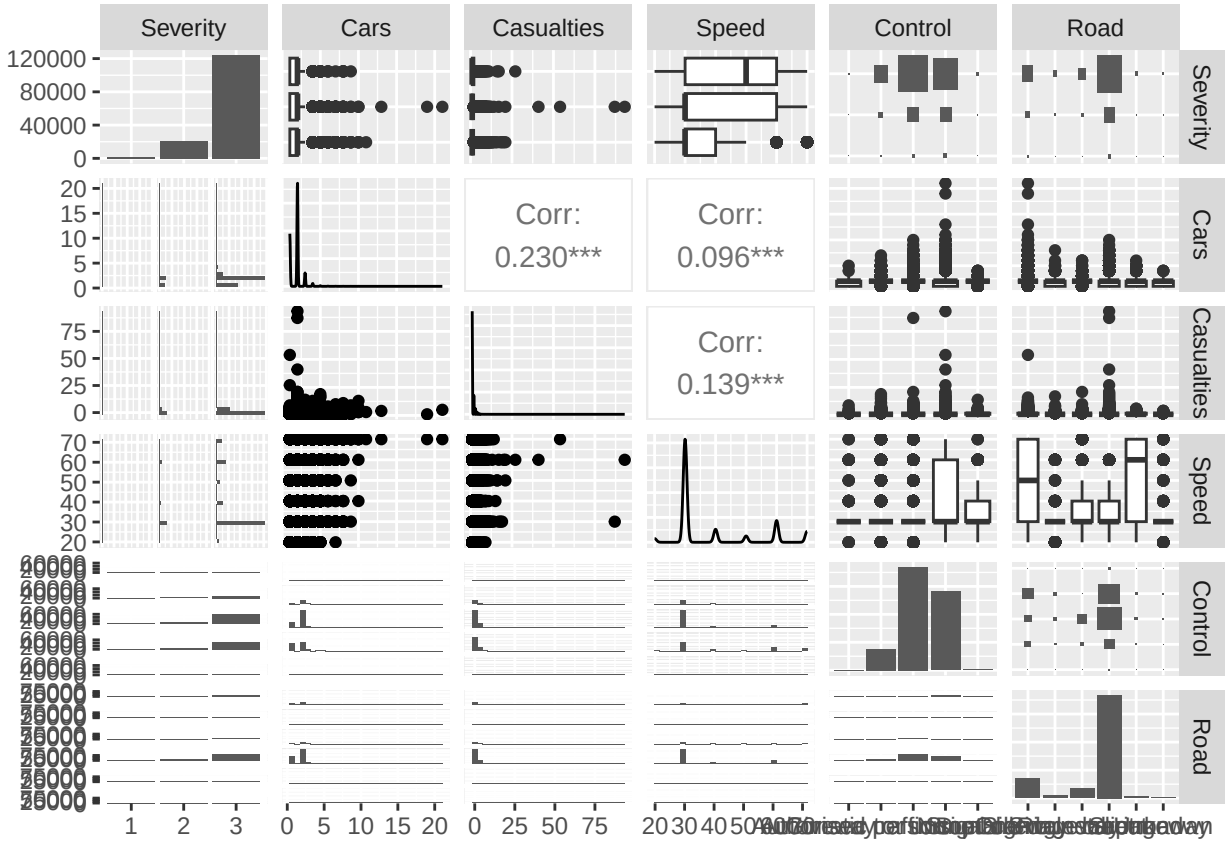
```
library(ggplot2)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.2.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
pm = ggpairs(subdata)
pm
```

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



As mentioned in previous task, the 6 variables of choice are Severity (accident severity levels from 1 to 3), Cars (number of cars involved in accident), Casualties (number of casualties in accident), Speed (speed-limit imposed at the location of the accident), Control (junction-control available at the location of the accident), and Road (type of road where accident occurred). Of these variables, Severity, Cars, Casualties, and Speed are continuous variables (but they are integer valued), and Control and Road are categorical variables.

In terms of relationships between variables, it appears no explanatory variables have very strong relationship with Severity. It appears that number of cars and casualties in an accident has relatively strong correlation, and it makes sense that if more cars are involved in an accident then more people are hurt. There is also relatively strong correlation between Speed and Casualties as well.

```
unique(subdata$Control)
```

```
## [1] None                      Giveway or uncontrolled  Automatic traffic signal
## [4] Authorised person         Stop Sign
## 5 Levels: Authorised person ... Stop Sign
```

```
unique(subdata$Road)
```

```
## [1] Single carriageway One way street     Roundabout         Dual carriageway
## [5] Unknown            Slip road
## 6 Levels: Dual carriageway One way street Roundabout ... Unknown
```

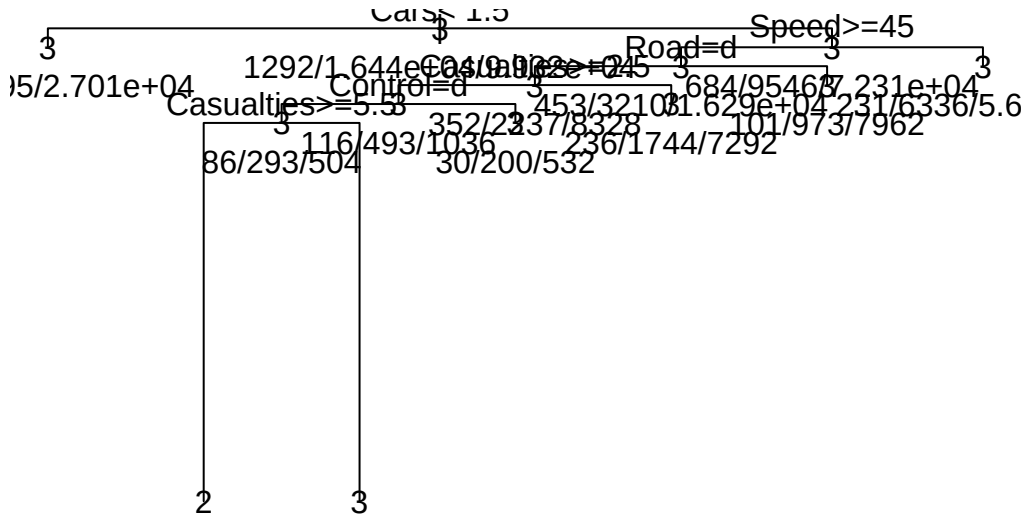
When looking into relationship between categorical variables, it appears that the number of cars and casualties involved in an accident is the highest when the junction-control in place is Giveway or None. This is an intuitive result as less traffic control means people are less cautious. Also, there seems to be higher number of cars involved if the accident takes place on a dual or single carriageway. This result is also intuitive since it is likely for larger accidents to happen on highways.

Observing the histogram of Severity, we can see that most observations have Severity level 3. This is highly skewed distribution of the response variable which may interfere with the validity of models in the following sections.

Task 4

```
library(rpart)
```

```
train_rows = sample(nrow(subdata), size = nrow(subdata) * 0.8)
train_data = subdata[train_rows,]
test_data = subdata[-train_rows,]
fit = rpart(Severity ~. , data = train_data, method = "class", cp=0.0001)
plot(fit)
text(fit, use.n=TRUE, all=TRUE, cex=1)
```




```
summary(fit)
```

```
## Call:
## rpart(formula = Severity ~ ., data = train_data, method = "class",
##     cp = 1e-04)
##   n= 117057
##
##             CP nsplit rel error    xerror        xstd
## 1 0.0001691761      0 1.0000000 1.0000000 0.006917311
## 2 0.0001000000      6 0.9989849 0.9994361 0.006915708
##
## Variable importance
##       Cars      Speed       Road Casualties    Control
##         47         28         16          8          2
##
## Node number 1: 117057 observations,    complexity param=0.0001691761
##   predicted class=3  expected loss=0.1514903  P(node) =1
##     class counts:  1292 16441 99324
##    probabilities: 0.011 0.140 0.849
##   left son=2 (34516 obs) right son=3 (82541 obs)
##   Primary splits:
##       Cars       < 1.5 to the left,  improve=386.79870, (0 missing)
##       Speed      < 45  to the right, improve=158.04180, (0 missing)
##       Control    splits as  RRRLR,   improve=124.69340, (0 missing)
##       Casualties < 2.5 to the right, improve= 58.44626, (0 missing)
```

```
##       Road       splits as  LLRLRL,  improve= 40.47967, (0 missing)
##
## Node number 2: 34516 observations
##   predicted class=3  expected loss=0.2173774  P(node) =0.2948649
##     class counts:   608  6895 27013
##    probabilities: 0.018 0.200 0.783
##
## Node number 3: 82541 observations,    complexity param=0.0001691761
##   predicted class=3  expected loss=0.1239384  P(node) =0.7051351
##     class counts:   684  9546 72311
##    probabilities: 0.008 0.116 0.876
##   left son=6 (19953 obs) right son=7 (62588 obs)
##   Primary splits:
##       Speed      < 45  to the right, improve=152.89900, (0 missing)
##       Casualties < 2.5 to the right, improve= 83.26241, (0 missing)
##       Control    splits as  RRRLR,   improve= 47.56831, (0 missing)
##       Road       splits as  RRRLRR,  improve= 27.97270, (0 missing)
##       Cars       < 4.5 to the right, improve= 11.19620, (0 missing)
##   Surrogate splits:
##       Road       splits as  LRRRLR,  agree=0.778, adj=0.080, (0 split)
##       Cars       < 4.5 to the right, agree=0.760, adj=0.006, (0 split)
##       Casualties < 7.5 to the right, agree=0.758, adj=0.000, (0 split)
##
## Node number 6: 19953 observations,    complexity param=0.0001691761
##   predicted class=3  expected loss=0.1835814  P(node) =0.1704554
##     class counts:   453  3210 16290
##    probabilities: 0.023 0.161 0.816
##   left son=12 (10917 obs) right son=13 (9036 obs)
##   Primary splits:
##       Road       splits as  RRRLRR,  improve=118.115500, (0 missing)
##       Casualties < 2.5 to the right, improve= 67.052760, (0 missing)
##       Speed      < 65  to the left,  improve= 52.848130, (0 missing)
##       Control    splits as  RRLLL,   improve=  8.515023, (0 missing)
##       Cars       < 4.5 to the right, improve=  4.277424, (0 missing)
##   Surrogate splits:
##       Speed      < 65  to the left,  agree=0.848, adj=0.665, (0 split)
##       Control    splits as  LRLLL,   agree=0.567, adj=0.045, (0 split)
##       Cars       < 2.5 to the left,  agree=0.565, adj=0.039, (0 split)
##       Casualties < 7.5 to the left,  agree=0.547, adj=0.001, (0 split)
##
## Node number 7: 62588 observations
##   predicted class=3  expected loss=0.1049243  P(node) =0.5346797
##     class counts:   231  6336 56021
##    probabilities: 0.004 0.101 0.895
##
## Node number 12: 10917 observations,    complexity param=0.0001691761
##   predicted class=3  expected loss=0.2371531  P(node) =0.09326226
##     class counts:   352  2237  8328
##    probabilities: 0.032 0.205 0.763
##   left son=24 (1645 obs) right son=25 (9272 obs)
##   Primary splits:
##       Casualties < 2.5 to the right, improve=54.5303900, (0 missing)
##       Control    splits as  RRRLR,   improve=12.4521300, (0 missing)
##       Cars       < 4.5 to the right, improve= 1.6748950, (0 missing)
```

```
##       Speed      < 55  to the right, improve= 0.5802951, (0 missing)
##   Surrogate splits:
##       Cars < 5.5 to the right, agree=0.85, adj=0.006, (0 split)
##
## Node number 13: 9036 observations
##   predicted class=3  expected loss=0.1188579  P(node) =0.07719316
##     class counts:   101   973  7962
##    probabilities: 0.011 0.108 0.881
##
## Node number 24: 1645 observations,    complexity param=0.0001691761
##   predicted class=3  expected loss=0.3702128  P(node) =0.01405298
##     class counts:   116   493  1036
##    probabilities: 0.071 0.300 0.630
##   left son=48 (883 obs) right son=49 (762 obs)
##   Primary splits:
##       Control    splits as  -RRLR,   improve=9.9815130, (0 missing)
##       Casualties < 5.5 to the right, improve=9.2606990, (0 missing)
##       Cars       < 3.5 to the left,  improve=1.3324730, (0 missing)
##       Speed      < 55  to the right, improve=0.2840629, (0 missing)
##   Surrogate splits:
##       Speed < 55  to the right, agree=0.548, adj=0.024, (0 split)
##
## Node number 25: 9272 observations
##   predicted class=3  expected loss=0.2135462  P(node) =0.07920927
##     class counts:   236  1744  7292
##    probabilities: 0.025 0.188 0.786
##
## Node number 48: 883 observations,    complexity param=0.0001691761
##   predicted class=3  expected loss=0.4292186  P(node) =0.007543334
##     class counts:    86   293   504
##    probabilities: 0.097 0.332 0.571
##   left son=96 (63 obs) right son=97 (820 obs)
##   Primary splits:
##       Casualties < 5.5 to the right, improve=9.9441690, (0 missing)
##       Cars       < 3.5 to the left,  improve=2.2485930, (0 missing)
##       Speed      < 55  to the right, improve=0.1085992, (0 missing)
##   Surrogate splits:
##       Cars < 6.5 to the right, agree=0.93, adj=0.016, (0 split)
##
## Node number 49: 762 observations
##   predicted class=3  expected loss=0.3018373  P(node) =0.006509649
##     class counts:    30   200   532
##    probabilities: 0.039 0.262 0.698
##
## Node number 96: 63 observations
##   predicted class=2  expected loss=0.4444444  P(node) =0.0005381993
##     class counts:    11    35    17
##    probabilities: 0.175 0.556 0.270
##
## Node number 97: 820 observations
##   predicted class=3  expected loss=0.4060976  P(node) =0.007005134
##     class counts:    75   258   487
##    probabilities: 0.091 0.315 0.594
```

The order of importance in the explanatory variables are Cars, Speed, Road, Casualties, and Control. One point of concern is that the distribution of Severity is highly skewed.

```
pred = predict(fit, test_data, type="class")
groundtruth = test_data$Severity
tab = table(pred, groundtruth)
tab
```

```
##      groundtruth
## pred     1     2     3
##    1     0     0     0
##    2     1     8     7
##    3   365  4227 24657
```

Since most of the accidents are classified as 3 on Severity level, it is highly likely that the mode Severity level in each leaf node of the classification tree is 3. Hence, this pushes the model to predict 3 almost all cases. Because there are relatively very few Severity level 1 accidents, the classification tree is not inclined to predict any observations as level 1.

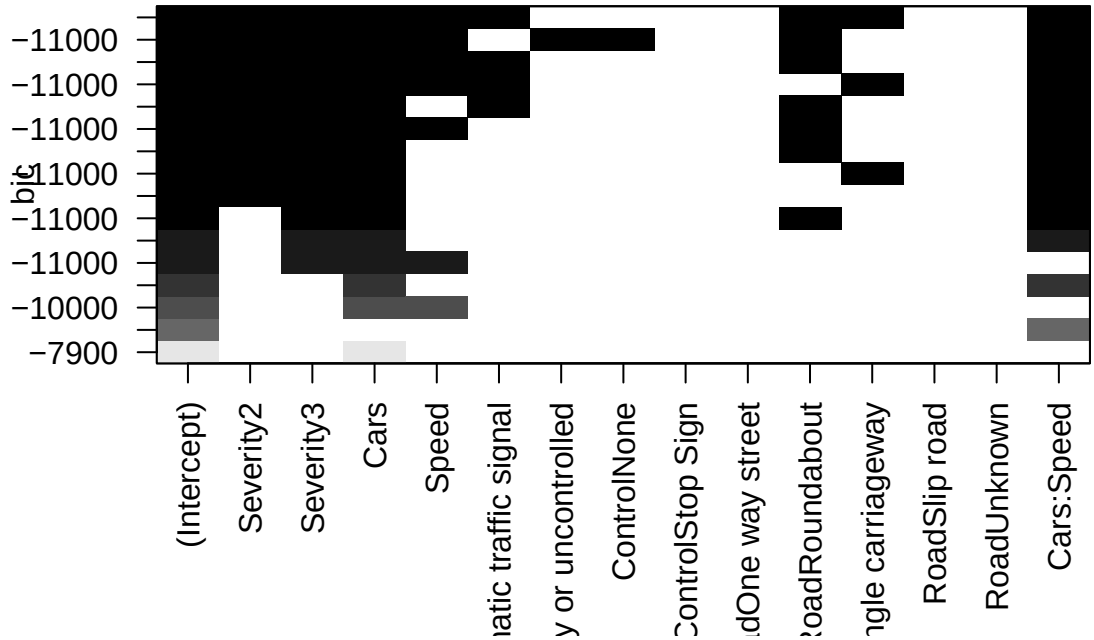
One simple prediction that can be made following the classification tree is on an accident involving 1 car. This is because the first node splits at Cars < 1.5, and any accidents satisfying this condition is classified into a leaf node with predicted Severity of 3. If an accident has 4 cars on a road where speed-limit was 40, it does not satisfy the first node condition Cars < 1.5 and it does not satisfy the following condition speed $>= 40$, hence it is assigned to the next leaf node with predicted Severity of 3.

Task 6

I will use best subsets for model selection. The continuous variable to predict will be Casualties. Among the explanatory variables, I will include the interaction between Cars and Speed variables.

```
library(leaps)
```

```
subsets = regsubsets(Casualties ~ Severity + Cars + Speed + Control + Road + Cars * Speed,
                     data = subdata, nbest = 2)
plot(subsets, scale="bic")
```



When selecting a model based on BIC, we choose the model with the lowest BIC. From the result of best subset plot, we can see that the best model consists of Severity, Cars, Speed and the interaction term. We see that one category from Control and two categories from Road are selected but not the rest. Since at least a part of each categorical variable is important in explaining Casualties, we should include both in the model.

```
model = lm(Casualties ~ Severity + Cars + Speed + Control + Road + Cars * Speed,
           data = subdata)
step_select = step(model, k=2)
```

```
## Start:  AIC=-56167.7
## Casualties ~ Severity + Cars + Speed + Control + Road + Cars *
##     Speed
##
##              Df Sum of Sq    RSS    AIC
## <none>                     99658 -56168
## - Control     4     63.99  99722 -56082
## - Cars:Speed  1     97.03  99755 -56027
## - Road        5    111.58  99769 -56014
## - Severity    2    450.16 100108 -55512
```

The result is consistent when performing stepwise selection, where the AIC is the lowest when no variable is removed. Hence, the model selection process suggests we should keep all 5 explanatory variables and the interaction term.

Task 8

Residual disclosure is a major concern for this dataset. The exact time and location of every accident in UK is available through longitude, latitude, date, and time variables. Especially for large scale accidents, it is likely that many media sources revealed the drivers or passengers involved. Hence, using this dataset and looking for past news articles that covered these accidents, personal information could be identified. Even if their personal information was anonymized on media, it is possible for a third party to identify the people involved in the accident. Assuming the people involved suffered injuries to their body or vehicle, it would be very easy to identify the people by anyone with access to some combination of hospital or insurance databases by matching the accident dates and locations.