NAÏVE BAYES USING R ON ACCIDENT DATASET

1. Explore the data: print summary statistics, plot distributions, and plot correlation plots. In complete sentences, describe what you notice/see from your exploration, noting things like potential outliers, variable types, descriptions of distributions (what are the skews), and identify significant relationships between variables, etc. Make sure to take note of any missing values and comment on any data cleaning you performed. Include any relevant plots in your report.

a. About the dataset

- The accidents Full dataset has 24 columns and 42,183 rows total. Most of the dataset is binary with a few numerical variables.
- There was no missing data, so no techniques were used to omit any data.
- An additional binary variable was added to the dataset to show whether an injury occurred or not.

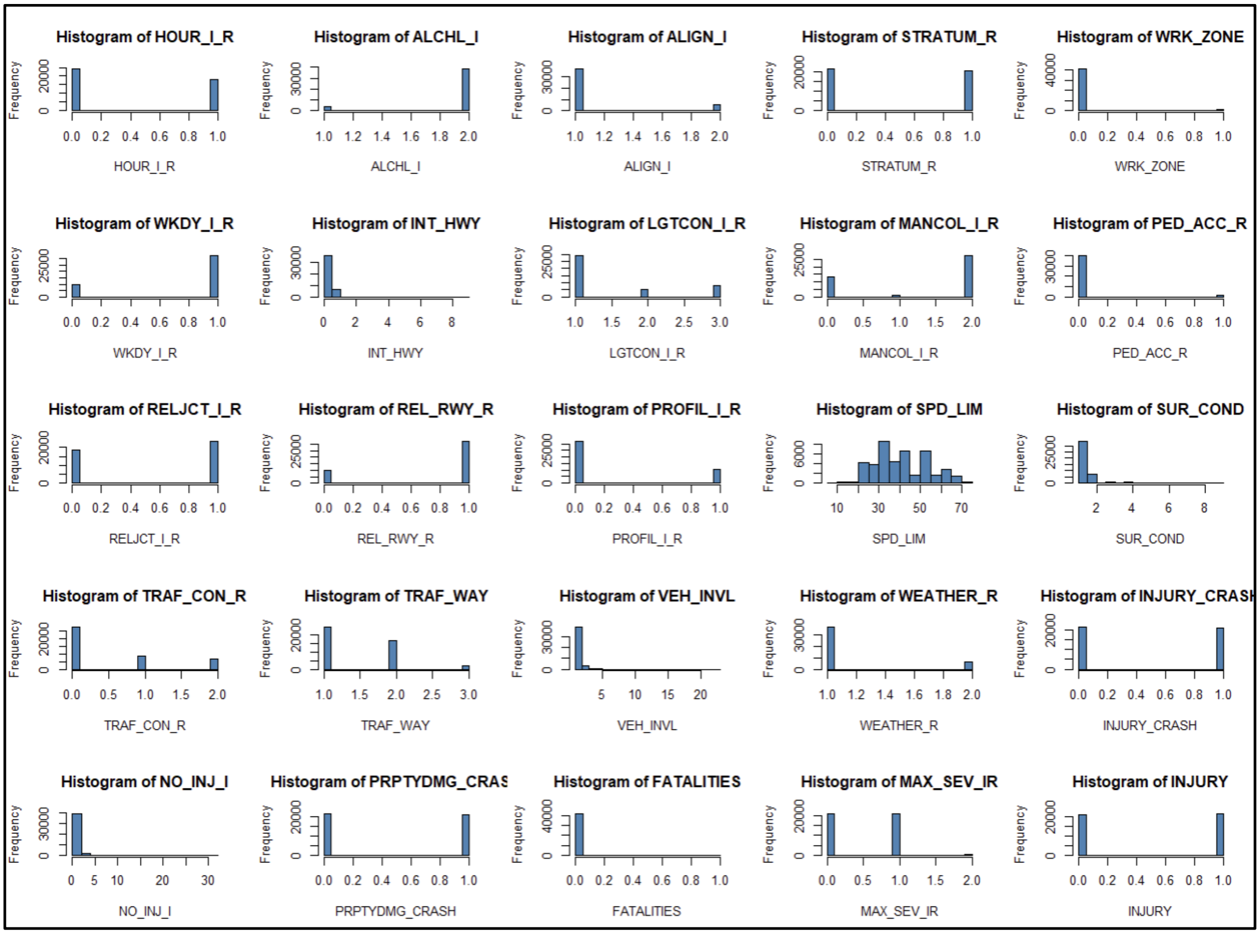
b. Overview of summary statistics

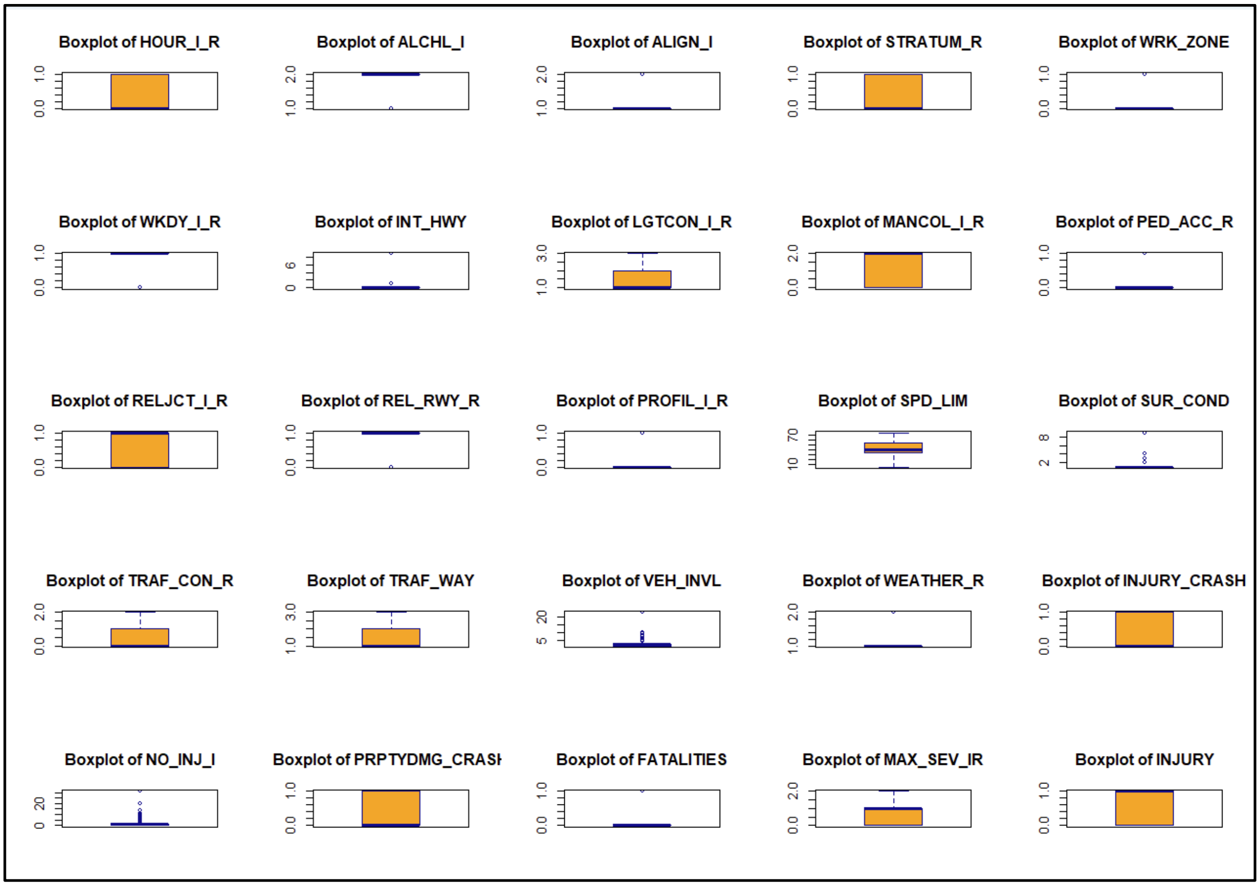
- SPD_LIM: The average speed limit at which an accident occurred was 43.55 mph.
- INJURY_CRASH: About half of the accidents caused an injury.

c. Visual Perspectives

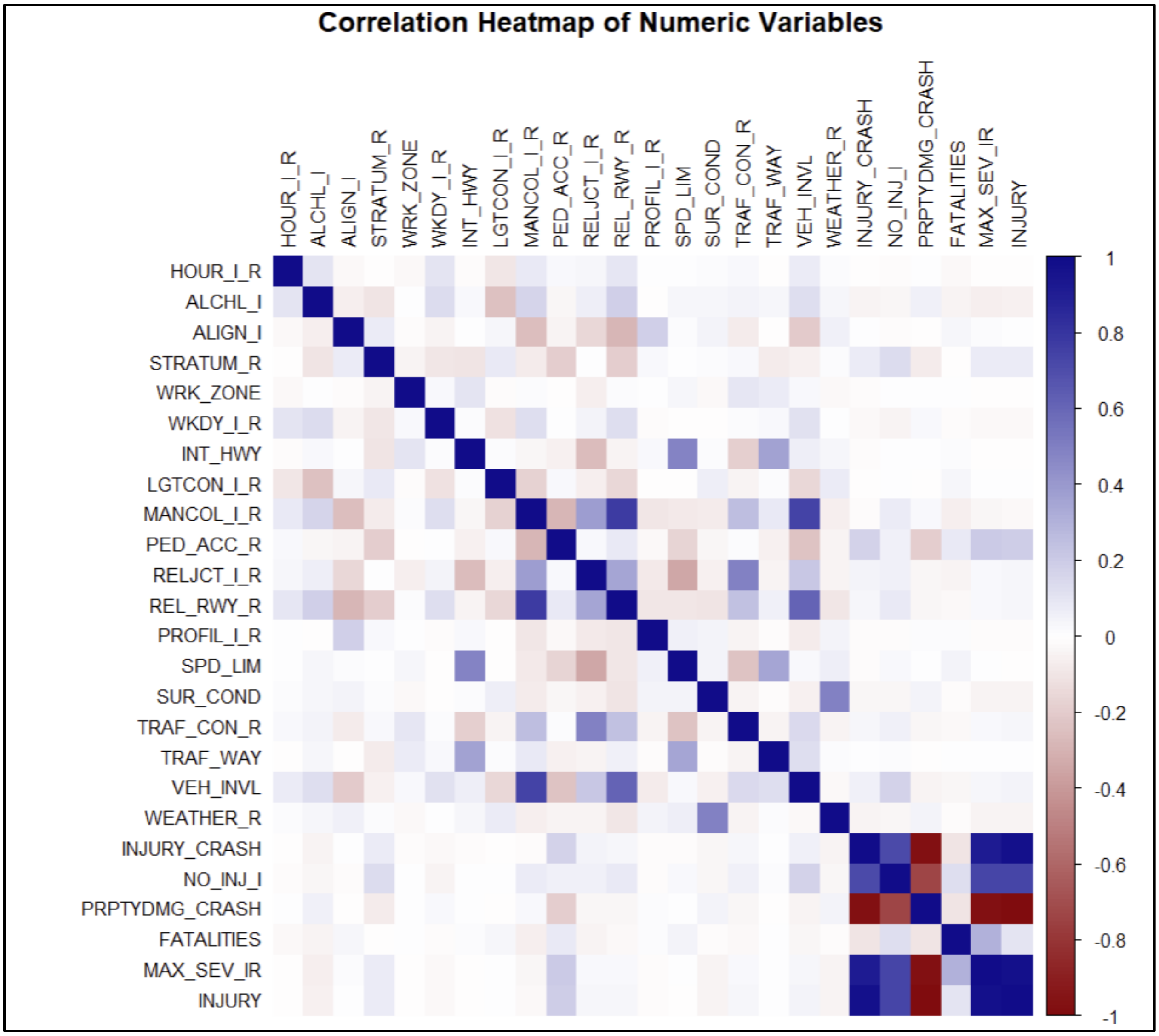
- Skews to note:
    - VEH_INVL: Right skewed
- Histograms:
    - INJURY_CRASH: About half of the accidents caused an injury
    - ALCHL_I: Alcohol was not involved in a majority of the accidents
    - SPD_LIM: It has bars across all the values
    - Many variables had, like, just two or three non-zero values, which are very large. Ex: No_INJ_I have a max value of 31 and a min value of 0. VEN_INVL has a min value of 1 and a max of 23.

```
   HOUR_I_R          ALCHL_I         ALIGN_I        STRATUM_R          WRK_ZONE
 Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.00000
 Median :0.0000   Median :2.000   Median :1.000   Median :0.0000   Median :0.00000
 Mean   :0.4293   Mean   :1.913   Mean   :1.132   Mean   :0.4916   Mean   :0.02262
 3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :2.000   Max.   :2.000   Max.   :1.0000   Max.   :1.00000
   WKDY_I_R          INT_HWY          LGTCON_I_R      MANCOL_I_R      PED_ACC_R
 Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   :0.000   Min.   :0.00000
 1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.00000
 Median :1.0000   Median :0.0000   Median :1.000   Median :2.000   Median :0.00000
 Mean   :0.7716   Mean   :0.1503   Mean   :1.493   Mean   :1.337   Mean   :0.04051
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :9.0000   Max.   :3.000   Max.   :2.000   Max.   :1.00000
  RELJCT_I_R        REL_RWY_R        PROFIL_I_R        SPD_LIM          SUR_COND
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 5.00   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:35.00   1st Qu.:1.000
 Median :1.0000   Median :1.0000   Median :0.0000   Median :40.00   Median :1.000
 Mean   :0.5579   Mean   :0.7665   Mean   :0.2432   Mean   :43.55   Mean   :1.291
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:55.00   3rd Qu.:1.000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :75.00   Max.   :9.000
  TRAF_CON_R        TRAF_WAY         VEH_INVL         WEATHER_R       INJURY_CRASH
 Min.   :0.0000   Min.   :1.000   Min.   : 1.000   Min.   :1.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:0.0000
 Median :0.0000   Median :1.000   Median : 2.000   Median :1.000   Median :0.0000
 Mean   :0.5163   Mean   :1.477   Mean   : 1.817   Mean   :1.143   Mean   :0.4977
 3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.: 2.000   3rd Qu.:1.000   3rd Qu.:1.0000
 Max.   :2.0000   Max.   :3.000   Max.   :23.000   Max.   :2.000   Max.   :1.0000
   NO_INJ_I       PRPTYDMG_CRASH     FATALITIES        MAX_SEV_IR         INJURY
 Min.   : 0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 1.0000   Median :0.0000   Median :0.00000   Median :1.0000   Median :1.0000
 Mean   : 0.7787   Mean   :0.4912   Mean   :0.01105   Mean   :0.5198   Mean   :0.5088
 3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :31.0000   Max.   :1.0000   Max.   :1.00000   Max.   :2.0000   Max.   :1.0000
```


Histogram of HOUR_I_R
Histogram of ALCHL_I
Histogram of ALIGN_I
Histogram of STRATUM_R
Histogram of WRK_ZONE
Histogram of WKDY_I_R
Histogram of INT_HWY
Histogram of LGTCON_I_R
Histogram of MANCOL_I_R
Histogram of PED_ACC_R
Histogram of RELJCT_I_R
Histogram of REL_RWY_R
Histogram of PROFIL_I_R
Histogram of SPD_LIM
Histogram of SUR_COND
Histogram of TRAF_CON_R
Histogram of TRAF_WAY
Histogram of VEH_INVL
Histogram of WEATHER_R
Histogram of INJURY_CRASH
Histogram of NO_INJ_I
Histogram of PRPTYDMG_CRAS
Histogram of FATALITIES
Histogram of MAX_SEV_IR
Histogram of INJURY



Boxplot of HOUR_I_R
Boxplot of ALCHL_I
Boxplot of ALIGN_I
Boxplot of STRATUM_R
Boxplot of WRK_ZONE
Boxplot of WKDY_I_R
Boxplot of INT_HWY
Boxplot of LGTCON_I_R
Boxplot of MANCOL_I_R
Boxplot of PED_ACC_R
Boxplot of RELJCT_I_R
Boxplot of REL_RWY_R
Boxplot of PROFIL_I_R
Boxplot of SPD_LIM
Boxplot of SUR_COND
Boxplot of TRAF_CON_R
Boxplot of TRAF_WAY
Boxplot of VEH_INVL
Boxplot of WEATHER_R
Boxplot of INJURY_CRASH
Boxplot of NO_INJ_I
Boxplot of PRPTYDMG_CRASH
Boxplot of FATALITIES
Boxplot of MAX_SEV_IR
Boxplot of INJURY



2. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why?

In this scenario, when no predictors or no further information are available, then we can make our predictions based on how often Injury happened, i.e., by seeing INJURY, which is our response variable. We can examine the frequency in the dataset to see if the occurrence of INJURY = Yes is more than that of INJURY = No. We can observe in the dataset that the number of times the Injury occurred (INJURY = Yes) is in approximately 51% of the cases; as compared to no injuries, this makes INJURY = Yes as the "majority class."

Therefore, in such situations when we have no context or predictors available to decide, the optimal strategy is to decide the most frequent outcome/class. In this case, we would predict INJURY= Yes, as it is the most occurring outcome or class in the dataset.

3.

a. Create a pivot table that examines INJURY as a function of the two predictors for these 12 records. Use all three variables in the pivot table as rows/columns.

	WEATHER	
TRAFFIC CON	1	2
0	2	1
1	0	0
2	0	0

The pivot table shows the injury counts across the combination of Weather_R and TRAF_CON_R for the 12 selected records. The table highlights that injuries occurred 2 times when TRAF_CON_R = 0 and Weather_R = 1. Injury occurred only once when TRAF_CON_R = 0 and Weather_R = 2. All other combinations show zero injuries.

b. Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.

Computing Probabilities using Bayes Theorem for 6 combinations

```
#P(Injury = 1 | WEATHER_R = 1, TRAF_CON_R = 0)
p1 <- 2/(2+1)
p1
#P(Injury = 1 | WEATHER_R = 1, TRAF_CON_R = 1)
p2 <- 0 / (0+1)
p2
#P(Injury = 1 | WEATHER_R = 1, TRAF_CON_R = 2)
p3 <- 0 / (0+1)
p3
#P(Injury = 1 | WEATHER_R = 2, TRAF_CON_R = 0)
p4 <- 1 / (1+5)
p4
#P(Injury = 1 | WEATHER_R = 2, TRAF_CON_R = 1)
p5 <- 0 / (0+1)
p5
#P(Injury = 1 | WEATHER_R = 2, TRAF_CON_R = 2)
p6 <- 0 / (0+0)
p6
```

c. Classify the 12 accidents using these probabilities and a cutoff of 0.5.

	WEATHER_R	TRAF_CON_R	prob	prediction
1	1	0	0.6666667	Yes
2	2	0	0.1666667	No
3	2	1	0.0000000	No
4	1	1	0.0000000	No
5	1	0	0.6666667	Yes
6	2	0	0.1666667	No
7	2	0	0.1666667	No
8	1	0	0.6666667	Yes
9	2	0	0.1666667	No
10	2	0	0.1666667	No
11	2	0	0.1666667	No
12	1	2	0.0000000	No

d. Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1

P(INJURY = 'Yes' | WEATHER_R = 1, TRAF_CON_R =1)
Calculating manually:
P(I='Yes'|W=1,T=1)= (P(I='Yes')*P(W=1| I = 'Yes')*P(T=1| I = 'Yes'))/(P(I='Yes')*P(W=1| I='Yes')*P(T=1| I='Yes'))+ (P(I='No')*P(W=1|I='No')*P(T=1|I='No'))

#P(I='Yes')
pa = 3/12

#P(W=1| I = 'Yes')
pb=2/3

#P(T=1| I = 'Yes')
pc=0/3 = 0

#P(I='No')
pd=9/12

#P(W=1|I='No')
pe=3/9

#P(T=1|I='No')
pf=2/9

Applying Bayes' theorem:

Probability = (pa*(pb*pc)/((pa*(pb*pc))+ (pd*(pe*pf))))

Probability = 0 as pc = 0

e. Run a naive Bayes classifier on the 12 records and two predictors using R or Orange. Check the model output to obtain probabilities and classifications for all 12 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent? Let us now return to the entire dataset.

```
INJURY WEATHER_R TRAF_CON_R    prob          prediction    prob_0
1       1         1          0 0.6666667          Yes    0.5000000
2       0         2          0 0.1666667           No    0.8000000
3       0         2          1 0.0000000           No    0.9992506
4       0         1          1 0.0000000           No    0.9970090
5       0         1          0 0.6666667          Yes    0.5000000
6       1         2          0 0.1666667           No    0.8000000
7       0         2          0 0.1666667           No    0.8000000
8       1         1          0 0.6666667          Yes    0.5000000
9       0         2          0 0.1666667           No    0.8000000
10      0         2          0 0.1666667           No    0.8000000
11      0         2          0 0.1666667           No    0.8000000
12      0         1          2 0.0000000           No    0.9940358
          prob_1    Predicted
1  0.5000000000         0
2  0.2000000000         0
3  0.0007494379         0
4  0.0029910269         0
5  0.5000000000         0
6  0.2000000000         0
7  0.2000000000         0
8  0.5000000000         0
9  0.2000000000         0
10 0.2000000000         0
11 0.2000000000         0
12 0.0059642147         0
```

No, the classification produced by the Naive Bayes model isn’t equivalent to that from the exact Bayes classification. We can see in the output that the exact Bayes approach classified 3 out of 12 accidents as causing Injury.

On the other hand, the Naive Bayes Classifier gave that no accident caused any injury. This is because the Naive Bayes Classifier classifies all 12 records as Injury = No, as the predicted probability of Injury = Yes was less than or equal to the cutoff value of 0.5.

In terms of ranking, the two methods differ. The exact Bayes uses the joint probabilities based on the two predictors WEATHER_R and TRAF_CON_R, while the Naive Bayes classifier assumes the conditional independence between these two predictors. These assumptions can distort the relative probabilities and lead to a different ordering of observations by injury risk.

Therefore, we can conclude that both the classification and the ranking of the observations are not equivalent between the two models even if they use same cutoff for classification.

4.

a. Assuming that no information or initial reports about the accident itself are available at the time of prediction (only location characteristics, weather conditions, etc.), which predictors can we include in the analysis? (Use the Data Dictionary below.)

We will include the following variables:

- SPD_LIM
- WRK_ZONE
- WEATHER_R
- TRAF_CON_R
- SUR_CON
- TRAF_WAY
- ALIGN_I
- PROFIL_I_R
- LGTCON_I_R
- INT_HWY
- REL_JCT_I_R
- REL_RWY_R
- WKDY_I_R
- HOUR_I_R

b. Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix.

Predicted	**Actual**	
	0	**1**
0	3659	3129
1	4708	5377

The above matrix shows that the model has correctly predicted 3659 accidents as No and 5377 accidents as causing Injury. However, we can see that it has misclassified 3129 accidents causing no injury and 4708 non-injuries as injuries.

c. What is the overall error for the validation set?

0.4644699
The Naive Bayes model achieved 46.45% on the dataset. In other words, the model correctly labels 53 out of every 100 accidents.

(5377 + 3659) / (5377 + 3659 + 4708 + 3129) = 9036 / 16873 = 53.55%

1-0.5355301 ~ 0.4645 = 46.45%

d. What is the percent improvement relative to the naive rule (using the validation set)?

6.334409
As the naive Bayes error is low as compared to the Naive rule error (used as a benchmark), we can say that our model performs better as compared to the Naive rule error. Percent improvement is approx 6%. This 6% improvement shows that the predictor variables provide meaningful predictions for predicting injury outcomes.

e. Examine the conditional probabilities output. Why do we get a probability of zero for P(INJURY = No | SPD_LIM = 5)?

The training dataset has no records where INJURY = No and SPD_LIM = 5. When the model is asked to calculate probabilities for this combination, it returns 0 because there are no observations to learn from. Since Naive Bayes multiplies conditional probabilities together, any zero probability makes the entire result zero. This is known as the Zero-Frequency Problem in Naive Bayes. This is why P(INJURY = No | SPD_LIM = 5) = 0. This issue can be addressed using Laplace Smoothing (also called additive smoothing), which adds a small constant to every category count. This ensures that every category has at least a small non-zero probability, making the model more robust and preventing zero probabilities from eliminating predictions entirely.