

---

# Data analysis on the CT Crash Dataset

---

Jianmin Chen, Jun Jin

December 4, 2019

## CONTENTS

<b>1</b>	<b>Abstract</b>	<b>2</b>
<b>2</b>	<b>Introduction and Description of Data</b>	<b>2</b>
2.1	Research proposing . . . . .	2
2.2	Qualitative an Quantitative . . . . .	2
<b>3</b>	<b>Exploratory Data Analysis</b>	<b>2</b>
3.1	Overall Distribution . . . . .	3
3.2	Crash vs Qualitative Variables . . . . .	5
3.2.1	Lighting . . . . .	5
3.2.2	Approach Left Turn and Approach Right Turn . . . . .	5
3.3	Crash vs Quantitative Variables . . . . .	6
3.3.1	AADT-MAJOR and AADT-MINOR . . . . .	6
3.3.2	Skew Angle . . . . .	7
3.4	EDA Conclusion . . . . .	8
<b>4</b>	<b>Statistical Inference</b>	<b>8</b>
4.1	One v.s. PDO crashing independence test . . . . .	8
4.2	One v.s. BC and KA crashing independence test . . . . .	11
4.3	Two v.s. crashing independence test . . . . .	12
4.3.1	Crossing type + skewed angle v.s. PDO crashing number . . . . .	12
4.3.2	Crossing type + Lighting vs Crash . . . . .	12
4.3.3	Crossing type + turns for Turn vs Crash . . . . .	13
4.3.4	Crossing type + AADT major for crash . . . . .	14
4.3.5	Crossing type + AADT minor for crash . . . . .	15
<b>5</b>	<b>Discussion of Results and Concluding Remarks</b>	<b>16</b>
5.1	Conclusions . . . . .	16
5.2	Further Discussion . . . . .	16
5.2.1	Logistic regression . . . . .	16
5.2.2	Sufficient dimension reduction . . . . .	16

<b>6 Appendix</b>	<b>18</b>
6.1 Extra Plots . . . . .	18
6.2 R codes . . . . .	22

## 1 ABSTRACT

In this paper, the group conducts the exploratory data analysis and statistical inferences on the CT crash data, judging that whether the variables such as type of intersections, presence of lighting, presence of a left-turn lane, presence of a right-turn lane and others could post great effects on the damage accounts and exploring that how they work. The research is useful for the government as a reference when designing the intersection in Connecticut.

## 2 INTRODUCTION AND DESCRIPTION OF DATA

In this paper, the group will take advantage of "CT Crash Data", it is the crash data for rural two-lane intersections and segments in Connecticut are collected in order to study highway safety. Generally, crash severity counts may be aggregated into three categories: PDO (property damage only), or B+C (possible injury or nonincapacitating injury), or K+A (fatal or incapacitating injury) crashes. Three different data sets are available on the HuskyCT course website, corresponding to the type of intersections. Dataset 3ST Intersections CT.csv consists of  $n = 385$  cases for 3ST (three-way stop controlled) intersections. Dataset 4ST Intersections CT.csv consists of  $n = 61$  cases for 4ST (four-way stop controlled) intersections. Dataset 4SG Intersections CT.csv consists of  $n = 102$  cases for 4SG (four-way signalized) intersections. Each dataset includes the following variables at each intersection: Intersection ID, Annual Average Daily Traffic (AADT) for major roads and minor roads, three dummy variables to indicate presence of lighting, presence of a left-turn lane, and presence of a right-turn lane, followed by skew angle, and then the counts of PDO, KA, and BC crashes.

### 2.1 Research proposing

In this paper, we are going to explore and confirm the following questions:

- Do these different types of intersection affect the number of crashes? (Regardless of other features.)
- If we control the types of intersection, whether other features (e.g. skewed angle, lighting, approaching turns, AADT major and minor) affect the number of crashes?
- And we may want to explore what kind of relationship (negative or positive) between the features we mentioned previously with the number of crashes.

### 2.2 Qualitative an Quantitative

Qualitative: Lighting, approach left, approach right.

Quantitative: AADT major, AADT minor, skew angle, PDO crashing number, KA crashing number, BC crashing number.

## 3 EXPLORATORY DATA ANALYSIS

We first rearranged the 3 data set and combined them into 1 single data set. One column is added to specify the intersection type, denoted as INTERTYPE. It has 3 values, "3ST", "4ST" and

"4SG". Then we have 548 observations in total.

A new categorical binary variable is constructed with "1" standing for there exists crash and "0" standing for no crash. This variable is denoted as "crash" and it is used as a candidate response variable when we want to analyze the data with mixed types of crash.

In this part, we will do exploratory data analysis to first find some insight of the data.

### 3.1 Overall Distribution

First, let us see overall distribution of the 3 types of crashes and the distribution across intersection types.

For the PDO(property damage only) crashes, around 38% of the intersections have 0 PDO crash overall. The overall distribution of PDO is spread out with some extreme values larger than 15. The distribution across different intersections shows some variations. The distribution is similar to the overall distribution in the 3ST type, while in 4SG and 4ST types, there are fewer zero counts and the data has larger variance with more non-zero counts had hence heavier tails. These infer that 4-way intersections tend to have more PDO crashes, especially the 4-way signalized intersection.

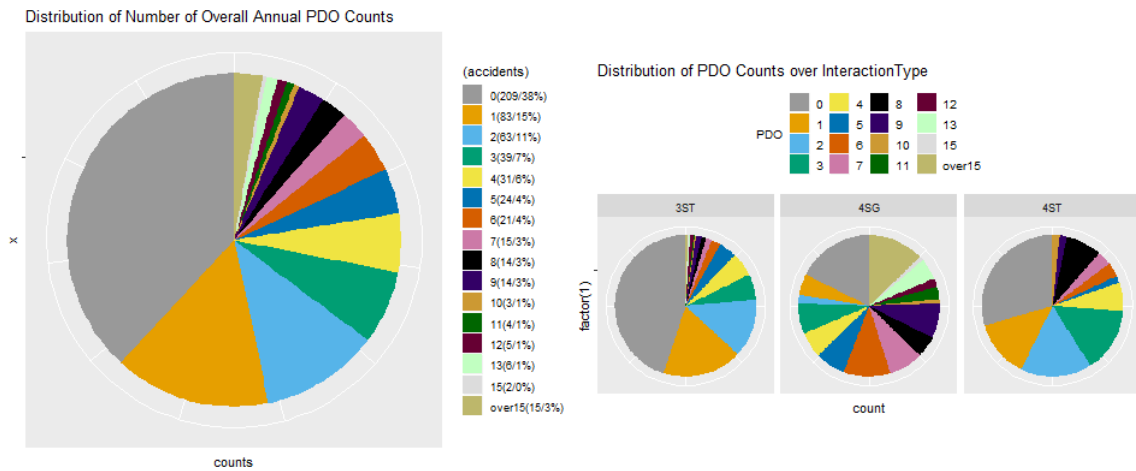


Figure 3.1: Pie chart of PDO

For the BC(possible injury or non-incapacitating injury) crashes, over one-half(59%) of the intersections have 0 crash and 76% of all intersections have less than 2 crashes. Like PDO, 4-way intersections tend to have larger number of BC crashes.

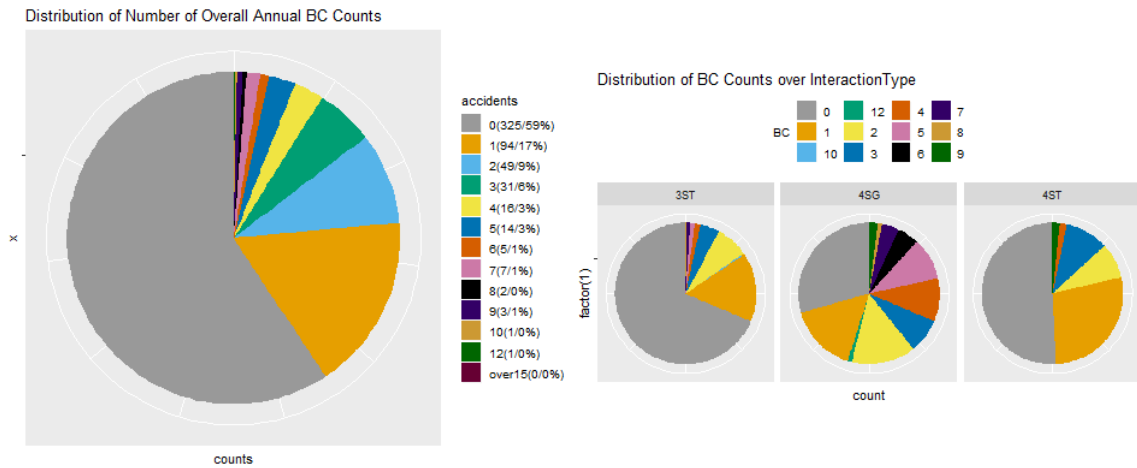


Figure 3.2: Pie chart of BC

For the KA(fatal or incapacitating injury) crashes, only 0,1 and 2 crashes are observed. We can infer that Poisson or Negative Binomial distribution may not be suitable to fit the KA data. Quite similar to the conclusion above, 4SG and 4ST types have larger proportion of intersection with crashes.

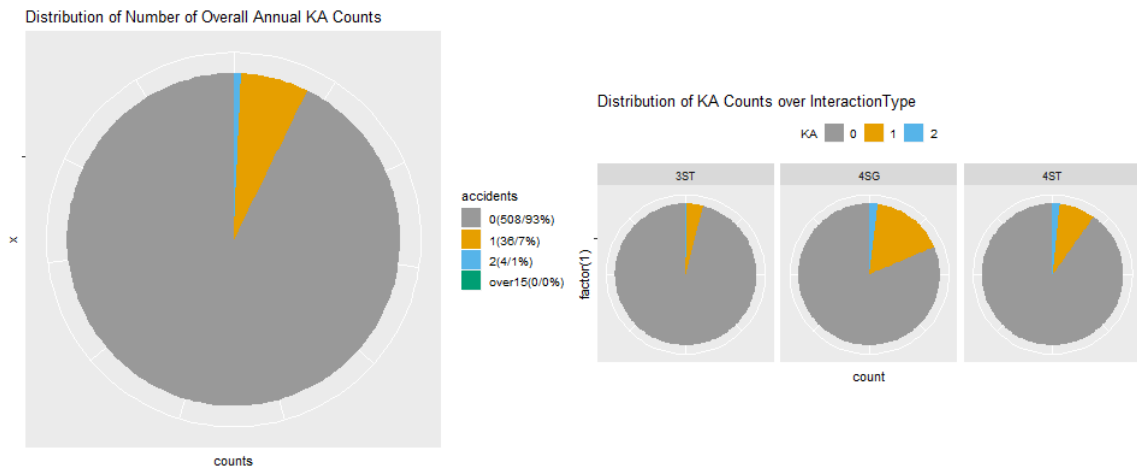


Figure 3.3: Pie chart of KA

Comparing the overall distribution of 3 types of crashes, we can see that for all 3 types, 0 crash has the maximum proportion in all 548 intersection. PDO and BC have more possible count values with larger dispersions, which is a hint that it is not plausible to mix PDO, KA and BC observations together when we want to model the distribution of crashes. Also, variance exists between types of intersections. Since intersection type is a more fixed factor that is hard to change in real traffic, in the following parts, we will analyze the data controlling intersection type to get some plausible suggestions to reduce crashes based on other variables. Since our aim is to find the vital factor of crashes, and KA crash can be considered as rare event, we will

also treat crash as binary distributed variables in some latter analysis, with 0 stands for no crash and 1 stands for at least 1 crash.

### 3.2 Crash vs Qualitative Variables

Next, we find some interesting relationship between crash and the 3 categorical variables, LIGHTING, APPROACH-LEFTTURN, and APPROACH-RIGHTTURN.

#### 3.2.1 Lighting

First, let's see the relationship between LIGHTING and crashes. There are 197 intersections without lighting in all 548(around 36%) intersections. Overall, when not controlling intersection type, intersection without lighting tends to have fewer crashes than intersection with lighting in all 3 types of crashes. We can compare the proportion of 0 counts in the following bar chart of PDO. The bar chart of BC and KA has similar tendency which are contained in appendix.

When controlling intersection type, there are some interesting findings.

For PDO, in 3ST and 4SG types, partitioned data has same tendency with whole data while for 4ST group, there is lower proportion of observations with crash when there is lighting, which is opposite to the whole data. For BC and KA, partitioning with intersection type generate same result in each group as the whole set.

The plots infer that there might be a relationship between lighting and crash. We will not take 4ST PDO group out for special consideration as these group has a small number of observations which may cause larger randomness.

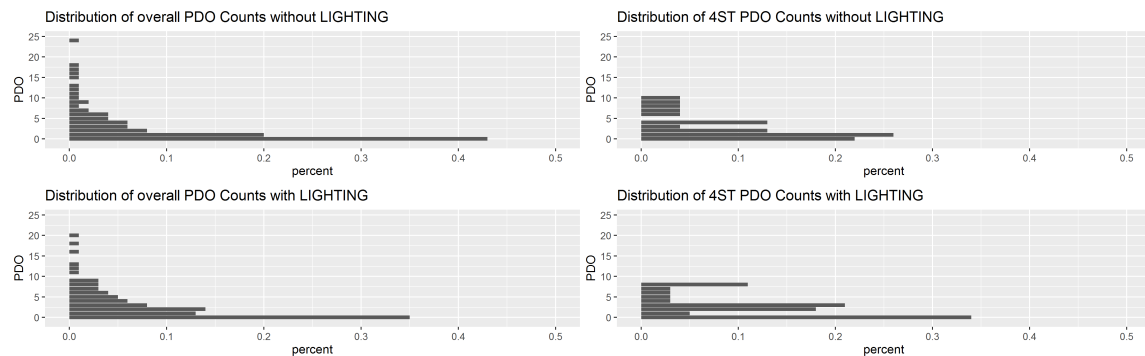


Figure 3.4: Bar Chart of Lighting

#### 3.2.2 Approach Left Turn and Approach Right Turn

There are only 65 intersections with approach left-turn. When partitioned with intersection type, we get the following contingency table. From the table we can see that the counts get even smaller when separated into 3 type, there is not much sense to analyze with data further separated by crash type. We discuss the relationship between Approach Left-Turn and crash based on the generated columns "crash".

There is similar problem with approach right-turn. And we also analyze based on crash as response variable. We observed similar association between these 2 variables and crash. Overall, when not controlling intersection type, intersection with approach left-turn or right turn has a higher crash rate, which infers that approach turn may has a positive relationship with car crashes. However, when we look into data grouped by intersection type, is it shown that left-turn or right turn does not have big effect with 3ST and 4SG intersections, even shown slightly

negative relationship. The effect is only obvious in 4ST groups, where the number of observations with left or right turn are quite small. Hence, the result of overall data and 4ST data may not be reliable. We need to check the relationship with data only from 3ST and 4SG group and take intersection type into consideration.

Further, as same patterns are observed with right-turn and left-turn, we combine these 2 variables into a new binary variable, denoted as "TURN", where "1" stands for there exists an approaching turn and "0" else. The follow graph is made with combined column. (Separated result is in Appendix.)

Table 3.1: Intersection vs Turns

Observations	3ST	4ST	4SG
Without Left-turn	368	58	57
With Left-turn	17	44	4

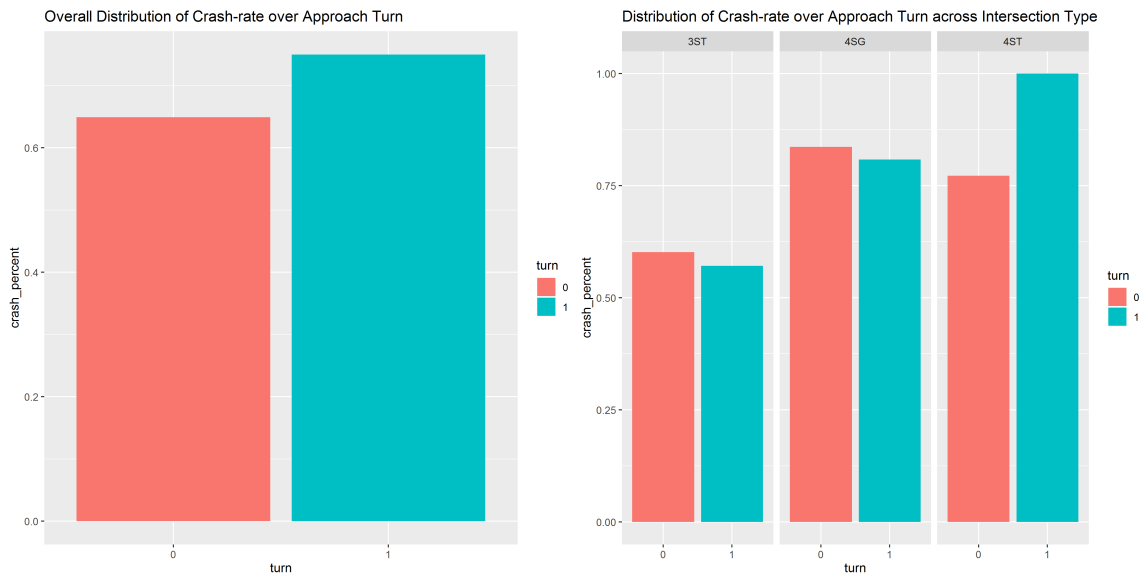


Figure 3.5: Bar Chart of Approach Left Turn

### 3.3 Crash vs Quantitative Variables

Also, there is likely to exist some association between AADT-MAJOR, AADT-MINOR, SKEW-ANGLE and crash.

#### 3.3.1 AADT-MAJOR and AADT-MINOR

Selected graphs are shown in this part.

We mainly apply the notched box-plot here to compare the distribution of AADT in intersection of crashes with intersection of no crash. The first 3 box plots below show significant difference on median of overall distribution as the notch does not overlap. This infers that generally, the intersections with crash has higher level of AADT-Major in all 3 types of crashes.

After controlling the intersection type, all partitioned data shows same pattern that higher AADT-Major appears in group with crash. However, for PDO and BC, the difference is only significant in 3ST group and for KA, no difference is significant.

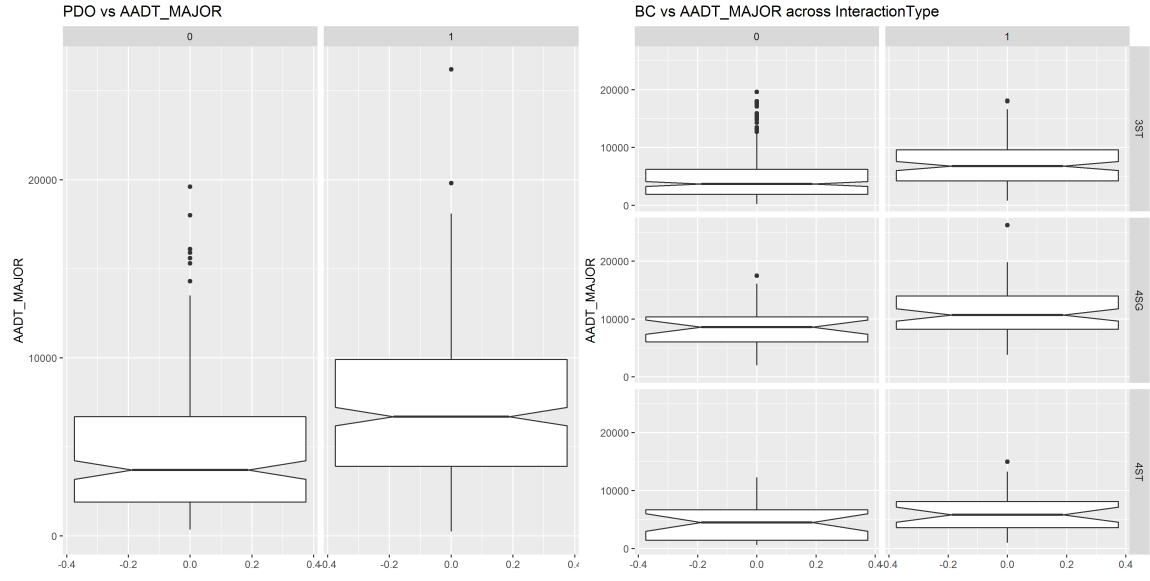


Figure 3.6: Boxplot of AADT-Major

For AADT-Minor, AADT-Minor is significantly larger in intersection with crash when exploring overall data for all of PDO, KA and BC. When taking intersection type into consideration, for 3ST group, same patterns are observed as overall data and the differences are all significant. However, opposite pattern occurs in 4SG group that intersection with crash has a lower AADT-MINOR value. The difference of sample median between the with-crash and without-crash group is significant in PDO crash. For 4ST group, all difference in median is not significant.

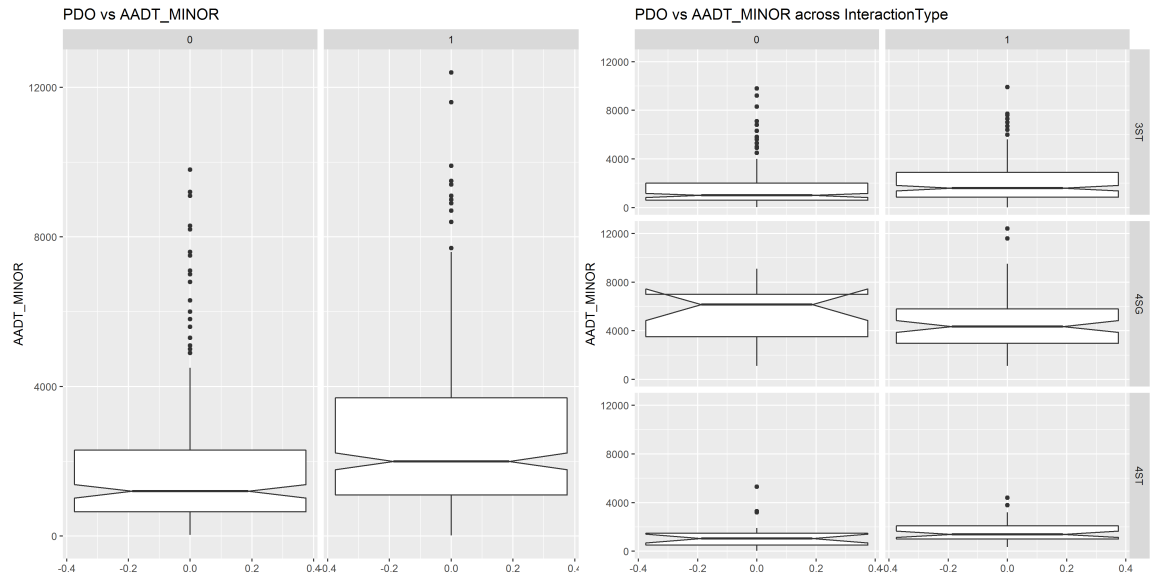


Figure 3.7: Boxplot of AADT-Minor

### 3.3.2 Skew Angle

For Skew Angle, we apply notched box plot also. In all 3 types of crash, groups with crash have smaller median for skew angle while the difference is only significant in KA crash type. When controlling intersection type, only difference in 3ST intersection and PDO crash shows significant larger pattern in no crash group. All others are insignificant. 4SG group gets insignificant

opposite pattern in PDO and BC crash. Skew angle is a variable that is considered less important.

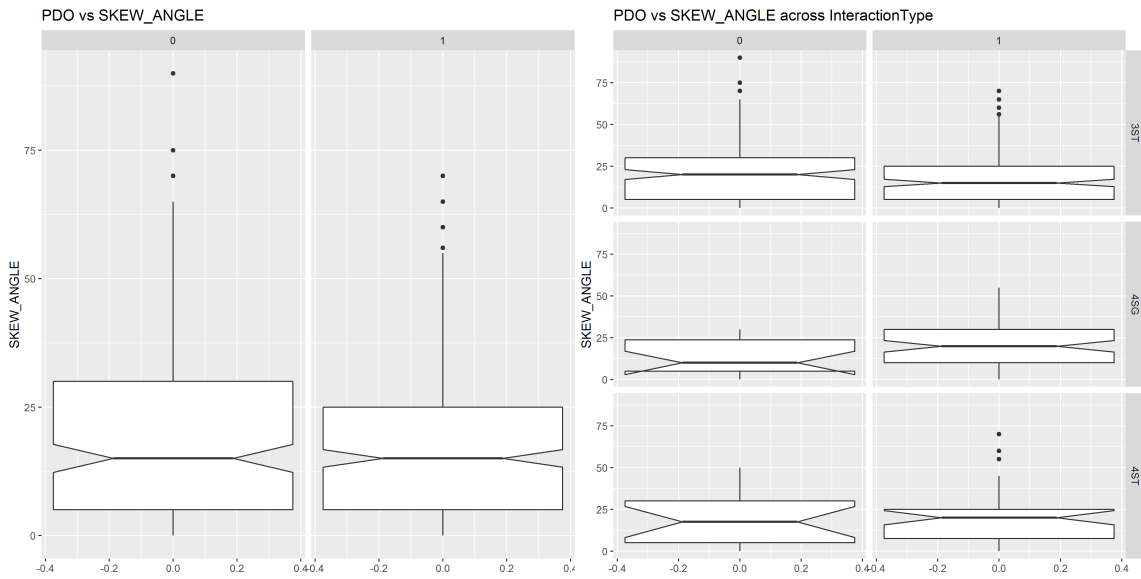


Figure 3.8: Boxplot of Skew Angle

### 3.4 EDA Conclusion

We get the following guess from the EDA part and the verification will be shown in the next inference part.

1. PDO, BC and KA has different distribution. They are not coming from different distribution family.
2. Intersection type is a vital factor to crash. Four-way intersections are at a higher probability for all 3 types of crash.
3. Intersection without lighting has a lower rate for all 3 types of crash. Controlling intersection type, same results are yielded except the PDO, 4ST group which has a lower crash rate when there is lighting. These results are kind of opposite to common sense.
4. Approach (left or right) turn has negative effect on overall crash rate, which infers that when there is an approach turn, the crash rate goes down. The result is only for 3ST and 4ST group individually. Data for 4SG group will not be analyzed with too few observations.
5. AADT-Major and AADT-Minor is at a higher level in intersections with crash for whole data set. When controlling intersection, However, for AADT-Major, the tendency is only significant in 3ST group with PDO and BC. For AADT-Minor, opposite significant pattern occurs in 4SG group of PDO crash.
6. Skew angle may not act as a vital factor to crash.

## 4 STATISTICAL INFERENCE

### 4.1 One v.s. PDO crashing independence test

We first test the relationship between crash and intersection type for 3 types of crash separately. We mainly use 3 types of methods here. When treating crash count as qualitative, normality assumption and same dispersion assumption are checked in each paired sample. Since none of the pairs satisfies normal assumption, we use Mann-Whitney test when paired sample has



same dispersion. When we get different dispersion, we calculate Bootstrap confidence interval for difference of sample mean to check whether the interval contains 0 or not. If the interval contains 0, we infer that there is not obvious difference on the sample mean. Re-sampling time is set as 1000 in Bootstrap method. The same dispersion tests are carried out by Ansari-Bradley nonparametric test here. Whether the impact is positive or negative can be directly got from the tests above.

When treating turning crash counts into binary variable as 1: crash, 0: nocrash, Pearson Chi-square test is applied to check independence.

For Intersection Type vs PDO crash, only the 4st vs 4sg group pass the same dispersion test. For 3st vs 4st and 3st vs 4sg, 95% Bootstrap Confidence Interval of difference of sample mean is calculate instead of testing. The result is shown in the table below.

Table 4.1: Bootstrap Confidence Interval for PDO crash counts mean difference

Difference of Mean	Lower Bound	Upper Bound	Contain 0
3ST-4ST	-1.652	-0.019	F(not obvious)
3ST-4SG	-6.670	-4.112	F

Firstly, we can see obviously that all PDO crashing number for three kinds of intersections are not normal data, so our test is based on Mann-Whitney test.

For PDO crashing between 3st intersections and 4st intersections,

$$H_0 : q_A = q_B$$

$$H_1 : q_A > q_B \text{ or } q_B > q_A$$

where  $q_A$  represents the median PDO crashing for 3st intersections,  $q_B$  represents the median PDO crashing for 4st intersections. Then, the statistic can be constructed by:

$$\frac{W - E(W)}{\sqrt{Var(W)}}$$

and

$$W = \sum_{j=1}^n R_j$$

$$E(W) = \frac{n(n+m+1)}{2}$$

$$Var(W) = \frac{mn}{12} \left\{ (n+m+1) - \frac{1}{(m+n)(m+n-1)} \sum_{j=1}^g t_j (t_j^2 - 1) \right\} \quad (4.1)$$

where  $n$  is PDO crashing sample size for 4st intersection;  $m$  is PDO crashing sample size for 3st intersection;  $R_j$  is the  $j$ -th sample rank for PDO crashing rank of 4st intersection samples;  $g$  is the tie group numbers and  $t_j$  is the tie group size for  $j$ -th tie group. According to the code in Appendix 2, our conclusion is  $q_B > q_A$ , i.e. the 4st intersection is more likely to happen PDO crash than 3st intersection.

For PDO crashing between 4st intersections and 4sg intersections,

$$H_0 : q_A = q_B$$

$$H_1 : q_A > q_B \text{ or } q_B > q_A$$

where  $q_A$  represents the median PDO crashing for 4st intersections,  $q_B$  represents the median PDO crashing for 4sg intersections. Then, the statistic can be constructed by:

$$\frac{W - E(W)}{\sqrt{Var(W)}}$$

and

$$W = \sum_{j=1}^n R_j$$

$$E(W) = \frac{n(n+m+1)}{2}$$

$$Var(W) = \frac{mn}{12} \left\{ (n+m+1) - \frac{1}{(m+n)(m+n-1)} \sum_{j=1}^g t_j (t_j^2 - 1) \right\}$$

where  $n$  is PDO crashing sample size for 4sg intersection;  $m$  is PDO crashing sample size for 4st intersection;  $R_j$  is the  $j$ -th sample rank for PDO crashing rank of 4sg intersection samples;  $g$  is the tie group numbers and  $t_j$  is the tie group size for  $j$ -th tie group. According to the code in Appendix 2, our conclusion is  $q_B > q_A$ , i.e. the 4sg intersection is more likely to happen PDO crash than 4st intersection.

For PDO crashing between 3st intersections and 4sg intersections,

$$H_0 : q_A = q_B$$

$$H_1 : q_A > q_B \text{ or } q_B > q_A$$

where  $q_A$  represents the median PDO crashing for 3st intersections,  $q_B$  represents the median PDO crashing for 4sg intersections. Then, the statistic can be constructed by:

$$\frac{W - E(W)}{\sqrt{Var(W)}}$$

and

$$W = \sum_{j=1}^n R_j$$

$$E(W) = \frac{n(n+m+1)}{2}$$

$$Var(W) = \frac{mn}{12} \left\{ (n+m+1) - \frac{1}{(m+n)(m+n-1)} \sum_{j=1}^g t_j (t_j^2 - 1) \right\}$$

where  $n$  is PDO crashing sample size for 4sg intersection;  $m$  is PDO crashing sample size for 3st intersection;  $R_j$  is the  $j$ -th sample rank for PDO crashing rank of 4sg intersection samples;  $g$  is the tie group numbers and  $t_j$  is the tie group size for  $j$ -th tie group. According to the code in Appendix 2, our conclusion is  $q_B > q_A$ , i.e. the 4sg intersection is more likely to happen PDO crash than 3st intersection.

According to the conclusions above, we have:

$$q_A < q_B < q_C$$

where  $q_A$  represents the median number of PDO crashes of 3st intersection,  $q_B$  represents the median number of PDO crashes of 4st intersection,  $q_C$  represents the median number of PDO crashes of 4sg intersection.

However, we can use another method - Pearson test in categorical analysis - to analyze this dependence. We firstly, transform the PDO crashing number into dummy variable. If the crashing number equals to 0, then we denote it as "noncrash", while if the crashing number is greater than 0, then we denote it as "crash". Thus, we can form the following table:

Table 4.2: Intersection VS Crash

	3ST	4ST	4SG	Total
Nocrash	173	18	18	209
Crash	212	43	84	339
Total	385	61	102	548

Then our hypothesis will be

$$H_0 : p_{ij} = p_i p_j$$

$$H_1 : p_{ij} \neq p_i p_j$$

and our statistic is:

$$Q_P = \frac{\sum_{i=1}^2 \sum_{j=1}^3 (n_{ij} - n_{i.} n_{.j} / n)^2}{n_{i.} n_{.j} / n}$$

and it approximately obeys a  $\chi^2$  (1) distribution. Thus, according to the code in Appendix 2, our conclusion is: they are dependence.

## 4.2 One v.s. BC and KA crashing independence test

We next conduct the independence test for BC crashes and KA crashes similar with the way we conduct for PDO test, so the testing procedures are omitted (codes are in Appendix 2). The conclusion is

- $q_A < q_B < q_C$ , where  $q_A$  represents the median number of BC crashes of 3st intersection,  $q_B$  represents the median number of BC crashes of 4st intersection,  $q_C$  represents the median number of BC crashes of 4sg intersection.
- $q_A < q_B = q_C$ , where  $q_A$  represents the median number of KA crashes of 3st intersection,  $q_B$  represents the median number of KA crashes of 4st intersection,  $q_C$  represents the median number of KA crashes of 4sg intersection.

and two categorical tables are:

Table 4.3: Categorical table for BC crash

	3ST	4ST	4SG	Total
Nocrash	264	31	30	325
Crash	121	30	72	223
Total	385	61	102	548

Table 4.4: Categorical table for KA crash

	3ST	4ST	4SG	Total
Nocrash	370	55	83	508
Crash	15	6	19	50
Total	385	61	102	548

And finally both of these two Pearson tests show dependences.

For 3 paired sample in KA crash, none of the pairs pass the Ansari-Bradley test. Hence, we construct Bootstrap Confidence Interval here.

Table 4.5: Bootstrap Confidence Interval for KA crash counts mean difference

Difference of Mean	Lower Bound	Upper Bound	Contain 0
3ST-4ST	-1.180	0.014	Y(not obvious)
3ST-4SG	-0.267	-0.078	F(not obvious)
4ST-4SG	-0.209	0.047	Y(not obvious)

For 3 paired sample in BC crash, none of the pairs pass the Ansari-Bradley test. Hence, we also construct Bootstrap Confidence Interval here.

Table 4.6: Bootstrap Confidence Interval for BC crash counts mean difference

Difference of Mean	Lower Bound	Upper Bound	Contain 0
3ST-4ST	-0.716	0.070	Y(not obvious)
3ST-4SG	-2.376	-1.361	F
4ST-4SG	-2.164	-0.930	F

From the discussion between intersection type and crash counts, we can draw the follow conclusion:

- (1) There are dependent between intersection type and crash counts for all 3 types of crashes by Pearson tests.
- (2) By Mann-whitney test, for PDO, 4SG intersection is more likely to happen PDO crash than 4ST intersection.
- (3) As there is still randomness in Bootstrap method, some intervals constructed above have bounds close to 0, which may change sign in other realizations. Hence, here we take  $|x| > 0.1$  as an obvious difference from 0. From the tables above, we can claim that for PDO, 4SG has obvious more crash than 3ST; for BC, 4SG has obvious more crash than 3ST and 4ST; for KA, there is no obvious pattern. However, all intervals constructed have the tendency to be at the left side of 0, which infers that there might be a tendency that the unit number of crash in all 3 types of crash follow the order  $3ST < 4ST < 4SG$ .

### 4.3 Two v.s. crashing independence test

#### 4.3.1 Crossing type + skewed angle v.s. PDO crashing number

As we can see that the skewed angle holds little choices of values, so we decide to transform skewed angle into categorical data and also transform PDO crashing number as our previous method, forming the categorical table below. We settle a standard: let  $s$  denote the skewed angle, if  $0 \leq s \leq 30$ , then we record it as “1”; If  $31 \leq s \leq 60$ , then we record it as “2”; If  $61 \leq s \leq 90$ , then we record it as “3”.

Table 4.7: Categorical table for PDO crash

	3ST_1	3ST_2	3ST_3	4ST_1	4ST_2	4ST_3	4SG_1	4SG_2	Total
Nocrash	131	31	11	14	4	0	18	0	209
Crash	168	39	5	36	6	1	71	13	339
Total	299	70	16	50	10	1	89	13	548

Then, the Pearson test shows that there exists a dependence (codes are in Appendix 2.)

#### 4.3.2 Crossing type + Lighting vs Crash

As the lighting itself is a dummy variable, we can form the following categorical table:

Table 4.8: Categorical table for PDO crash

	3ST_0	3ST_1	4ST_0	4ST_1	4SG_0	4SG_1	Total
Nocrash	74	99	5	13	6	12	209
Crash	76	136	18	25	28	66	339
Total	150	235	23	38	24	78	548

Then, the Pearson test shows that there exists a dependence (codes are in Appendix 2.) Here we get the 3 2-by-2 contingency table between Lighting and crash with PDO crash, controlling intersection type. Here, '0' stands for no Lighting, '1' stands for Lighting.

Table 4.9: Categorical table for PDO crash

	3ST_0	3ST_1	Total	4ST_0	4ST_1	Total	4SG_0	4SG_1	Total
Nocrash	74	99	173	5	13	18	6	12	18
Crash	76	136	212	18	25	43	28	66	94
Total	150	235	385	23	38	61	24	78	102

Same contingency tables are constructed be BC crash, KA crash and mixed crash type. Only table of KA, 4ST has cell counts less than 5. For this one, we apply Fisher-exact test. For others, we conduct individual Pearson test on each table. None of the above tests reject the Null hypothesis. P-value are shown in table below. When testing the overall effect across table, in the Mantel-Haenszel test cannot reject the Null in all 3 types of crashes. Hence, no dependence of Lighting and crash are observed here.

When we mix 3 types of crash together and still construct 2-by-2 table across intersections, same results are generated.

In all, we conclude that there is no effect of Lighting on crash when controlling intersection type.

Table 4.10: P-value for Lighting vs Crash

p-value	3ST	4ST	4SG
PDO	0.1657	0.3006	0.7644
KA	0.6484	0.0746	0.7509
BC	0.247	0.4883	0.976
Mixed	0.0879	0.9494	0.5732

#### 4.3.3 Crossing type + turns for Turn vs Crash

As the lighting itself is a dummy variable, we can also form the following categorical table:

Table 4.11: Categorical table for PDO crash

	3ST_00	3ST_01	3ST_10	3ST_11	4ST_00	4ST_10	4ST_11	4SG_00	4SG_01	4SG_10	4SG_11	Total
Nocrash	164	2	3	4	18	0	0	9	1	4	4	209
Crash	200	2	7	3	39	3	1	46	2	18	18	339
Total	364	4	10	7	57	3	1	55	3	22	22	548

where "00" represents no left approach or right approach; "01" represents no left approach but a right approach; "10" represents no right approach but left approach; "11" represents a left approach and right approach. Then, the Pearson test shows that there exists a dependence (codes are in Appendix 2.)

Based on the results from the EDA part, we combine Approach Left Turn and Approach Right Turn as one dummy variable where 1 stands for there is an approach turn and 0 stands for not. Even then, approach turn existence is still a rare event in all intersection types. Also, if we continuing partitioning the data by crash type to form contingency table, we will have cells with extremely small counts equal to 0. Hence, we only analyses with mixed data with all crashes types mixed. We can form the following categorical table:

Table 4.12: Categorical table for Turn vs Crash

counts	3ST_0	3ST_1	Total	4ST_0	4ST_1	Total	4SG_0	4SG_1	Total
Nocrash	145	9	154	13	0	13	9	9	18
Crash	219	12	231	44	4	48	46	38	84
Total	364	21	385	57	4	61	55	47	102
Test	Pearson	p-value	0.7834	Fisher	p-value	0.5691	Pearson	p-value	0.713

We do both Mantel-Haenszel test and 3 individual tests to test for independence between column and row.

For the 3 individual test, for 3ST and 4SG, we have all cell counts larger than 5. We apply Pearson Chi-square tests and get p-value as 0.7834 and 0.713, which infers there is no dependency. For 4ST, we use the Fisher-exact test as counts are rather small and get p-value of 0.3729. The individual tests show no relationship between Crash and Turns.

For the Mantel-Haenszel test, although we get the result that there is no overall effect of Turn on crash when controlling intersection type, the assumptions are not all satisfied due to small cell counts. But, it can still be some kind of evidence that there is no relationship.

#### 4.3.4 Crossing type + AADT major for crash

We firstly partition the full samples into 6 parts: 3st intersection + crash; 3st intersection + noncrash; 4st intersection + crash; 4st intersection + noncrash; 4sg intersection + crash; 4sg intersection + noncrash. And we carry out a Shapiro normality test on these 6 sample pieces, however, few of them pass the test. Hence, once again, we need to turn to the Mann-Whitney test.

Comparing the AADT major and controlling the intersection type in 3st, we have the following hypotheses:

$$H_0 : q_A = q_B$$

$$H_1 : q_A < q_B \text{ or } q_A > q_B$$

where  $q_A$  represents the median of AADT major in PDO noncrash situation in 3st intersection, while  $q_B$  represents the median of AADT major with PDO crash situation in 3st intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A < q_B.$$

Comparing the AADT major and controlling the intersection type in 4st, we have the following hypotheses:

$$H_0 : q_A = q_B$$

$$H_1 : q_A < q_B \text{ or } q_A > q_B$$

where  $q_A$  represents the median of AADT major in PDO noncrash situation in 4st intersection, while  $q_B$  represents the median of AADT major with PDO crash situation in 4st intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A = q_B.$$

Comparing the AADT major and controlling the intersection type in 4sg, we have the following hypotheses:

$$\begin{aligned} H_0 : q_A &= q_B \\ H_1 : q_A &< q_B \text{ or } q_A > q_B \end{aligned}$$

where  $q_A$  represents the median of AADT major in PDO noncrash situation in 4sg intersection, while  $q_B$  represents the median of AADT major with PDO crash situation in 4sg intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A = q_B.$$

By same Mann-Whitney tests, for KA crash, AADT-Major only has significant effect at 3ST intersection. Larger AADT-Major intersection has higher crash rate. For BC crash, significant impact is observed both in 3ST and 4ST. In both case, larger AADT-Major is related to higher crash rate.

#### 4.3.5 Crossing type + AADT minor for crash

With the same partition made for AADT major, we also carry out a Shapiro normality test on these 6 sample pieces for AADT minor, however, few of them pass the test. Hence, once again, we need to turn to the Mann-Whitney test.

Comparing the AADT minor and controlling the intersection type in 3st, we have the following hypotheses:

$$\begin{aligned} H_0 : q_A &= q_B \\ H_1 : q_A &< q_B \text{ or } q_A > q_B \end{aligned}$$

where  $q_A$  represents the median of AADT minor in PDO noncrash situation in 3st intersection, while  $q_B$  represents the median of AADT minor with PDO crash situation in 3st intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A < q_B.$$

Comparing the AADT minor and controlling the intersection type in 4st, we have the following hypotheses:

$$\begin{aligned} H_0 : q_A &= q_B \\ H_1 : q_A &< q_B \text{ or } q_A > q_B \end{aligned}$$

where  $q_A$  represents the median of AADT minor in PDO noncrash situation in 4st intersection, while  $q_B$  represents the median of AADT minor with PDO crash situation in 4st intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A = q_B.$$

Comparing the AADT minor and controlling the intersection type in 4sg, we have the following hypotheses:

$$\begin{aligned} H_0 : q_A &= q_B \\ H_1 : q_A &< q_B \text{ or } q_A > q_B \end{aligned}$$

where  $q_A$  represents the median of AADT minor in PDO noncrash situation in 4sg intersection, while  $q_B$  represents the median of AADT minor with PDO crash situation in 4sg intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A = q_B.$$

By same Mann-Whitney tests, for KA crash, no significant effects is observed. For BC crash, positive effect on crash rate is observed in 3ST intersection. For 4ST group, the AADT-Minor in crash and noncrash group do not pass the same dispersion test. We again carry out the Bootstrap Confidence Interval method on the difference of sample mean. The 95% interval for  $\text{mean}(\text{crash}) - \text{mean}(\text{noncrash})$  is (100,1051.25). Hence, there is obvious positive effect of AADT-Minor on crash rate in 4ST.

## 5 DISCUSSION OF RESULTS AND CONCLUDING REMARKS

### 5.1 Conclusions

From all the discussion above, we can draw the conclusion below:

For Intersection Type vs crash counts only: Intersection type affects the probability of all 3 types of crash. There is a tendency that the crash probability in 3 types of intersection follows that order  $3ST < 4ST < 4SG$ . The tendency is obvious in PDO, 4SG vs 4ST and BC, 3ST vs 4ST vs 4SG. We may recommend to pay much more attention when coming across a 4SG intersection.

Controlling intersection type:

- (1) Lighting: Discussions only focus on data with 3 types of crashes mixed together due to lack of observations. There is no significant effect of Lighting on crash rate observed.
- (2) Turns: Discussions only focus on data with 3 types of crashes mixed together due to lack of observations. There is no significant effect of Approach Turns on crash rate observed.
- (3) AADT-Major and AADT-Minor: Only positive impact of AADT-Major and AADT-Minor is observed on crash rate. For PDO, 3ST intersection with higher AADT-Major level or higher AADT-Minor level tends to has higher crash rate. For KA, same impact observed in 3ST. For BC, such positive effect occurs in 3ST and 4ST. The result infers that controlling AADT at a lower level might be a good method to reduce crash rate at 3ST and 4ST intersection, but might not help for 4SG intersections.

### 5.2 Further Discussion

For all the analysis above, we mainly apply 3 types of technique: data visualization, confidence interval construction and hypothesis testing on location and scale. Here are some other probable methods that can be applied to explore the data set.

#### 5.2.1 Logistic regression

All the methods used can hardly predict the crash rate or crash counts. Also, the impact of certain variable when controlling 2 or more other variables cannot be discusses with current technique. Since we have counts as response variable and also many 0 in the data set, logistic models are a good option here to fit the data, either with normal family(predict crash rate) or with poisson/negative-binomial family(predict crash counts). Then we can get prediction of crash and also further explore the effect of regressors in the full model.

#### 5.2.2 Sufficient dimension reduction

In statistical fields, there is a special topic - sufficient dimension reduction - which is used to select most relative variable to the response. Mathematically, for  $Y$  as response and  $X$  as a  $p$ -dimension predictor, we'd like to find a  $p \times d$  ( $d < p$ ) vector  $\eta$  such that

$$Y \perp X | \eta^T X.$$



where  $\eta^T X$  is denoted as the central subspace  $S_{Y|X}$ . Proposed by Li et. al.[1], the principal support vector machines is a popular sufficient dimension reduction method. The estimation procedure can be divided into three steps:

- 1. Compute the sample mean  $\bar{X}$  and sample variance matrix  $\hat{\Sigma}$ .
- 2. (Optimization) Let  $q_r, r = 1, \dots, h-1$  be  $h-1$  dividing points, equally spaced sample percentiles of  $\{Y_1, \dots, Y_n\}$ . Let

$$\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r).$$

Then let  $(\hat{\beta}_r, \hat{t}_r)$  be the minimizer of

$$\beta^T \hat{\Sigma} \beta + \lambda E_n \left\{ 1 - \tilde{Y}^r \left[ (X - \bar{X})^T \beta - t \right] \right\}^+.$$

- 3. (Estimation) Let  $\hat{v}_1, \dots, \hat{v}_d$  be the  $d$  leading eigenvectors of matrix

$$\hat{M}_n = \sum_{r=1}^{h-1} \hat{\beta}_r \hat{\beta}_r^T.$$

Then we use subspace spanned by  $\hat{v} = (\hat{v}_1, \dots, \hat{v}_d)$  to estimate  $S_{Y|X}$ .

By the way, proposed by Xia et. al.[2], the eigenvector number  $d$  can be selected by the jump ratio of eigenvalues. That is, if  $\lambda_1, \dots, \lambda_p$  is  $p$  eigenvalues of the matrix in descending order, then we forming the  $p-1$  jump ratio by

$$\{J_1, \dots, J_{p-1}\}, J_i = \frac{\lambda_i}{\lambda_{i+1}},$$

then we select  $d$  as

$$d = \arg \max_i J_i.$$

Geometrically, the following figure shows the principal.

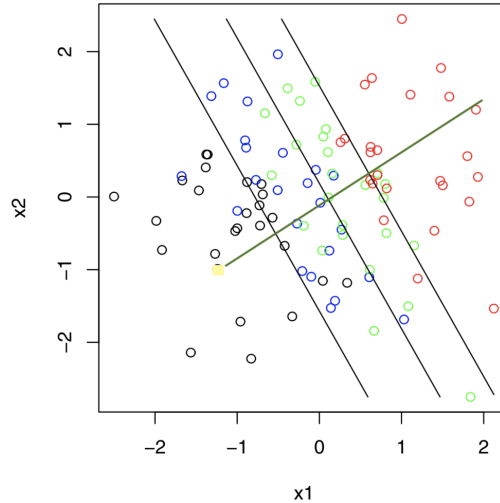


Figure 5.1: 2 dimension SVM surfaces and their normal line

In the figure, if the colors represent the value, from red (4th quantile part) to black (1st quantile part), the value  $Y$  descends. Then, three black surface is the SVM boundary for its four quantile

parts. Finally, the green line - the normal direction of its SVM boundary - can perfectly delineate a path (forming by the weights of  $x_1$  and  $x_2$ ) where the  $Y$  value grows fastest.

Implementing this algorithm by Python (code is in Appendix 2), we get the central subspace for PDO crashing number as response:

```
array([[ 0.00000000e+00, -6.28336625e-03, -3.44987939e-04, -1.54542490e-04],
       [ 9.99807242e-01, -1.96296091e-02, -3.32648777e-04, -1.63439760e-04],
       [-9.52640689e-04, -2.96027569e-02, -9.24611159e-01, -3.78741274e-01],
       [-1.67939885e-04, -7.34446402e-03, -2.37055602e-01, 6.33323237e-01],
       [-7.19757840e-05, -6.52082837e-03, -2.96534802e-01, 6.74871829e-01],
       [ 1.96096296e-02,  9.99300962e-01, -3.10761591e-02, -2.16533571e-03]])
```

It shows that for PDO crashing number, there are totally 4 factors which may significantly affects it. The AADT major holds a positive effect ( 9.99807242e-01) to PDO crash; The skewed angle holds a positive effect (9.99300962e-01) to PDO crash; The lighting holds a negative effect (-9.24611159e-01) to PDO crash; The right-turn approach holds positive effect (6.74871829e-01) to PDO crash.

Similarly, for BC crashing number as response, the central subspace will be:

```
array([[ 4.00994989e-05,  1.88638392e-05,  2.29261861e-04],
       [-2.92592349e-05, -2.48004596e-06, -4.20608685e-05],
       [-9.87266764e-01, -1.21357177e-02, -1.58483896e-01],
       [-1.53212888e-01,  3.38178083e-01,  9.28526411e-01],
       [ 4.23293850e-02,  9.41003883e-01, -3.35737815e-01],
       [-6.19423106e-03,  1.07378434e-05, -1.27455844e-03]])
```

It can also reveal three key and relative factor: lighting for negative effect, left-turn and right-turn approaches for positive effect. And certainly we can do the same procedure on the KA crashing easily, so we omit it here.

## 6 APPENDIX

### 6.1 Extra Plots

Extra support plots for AADT-Major:

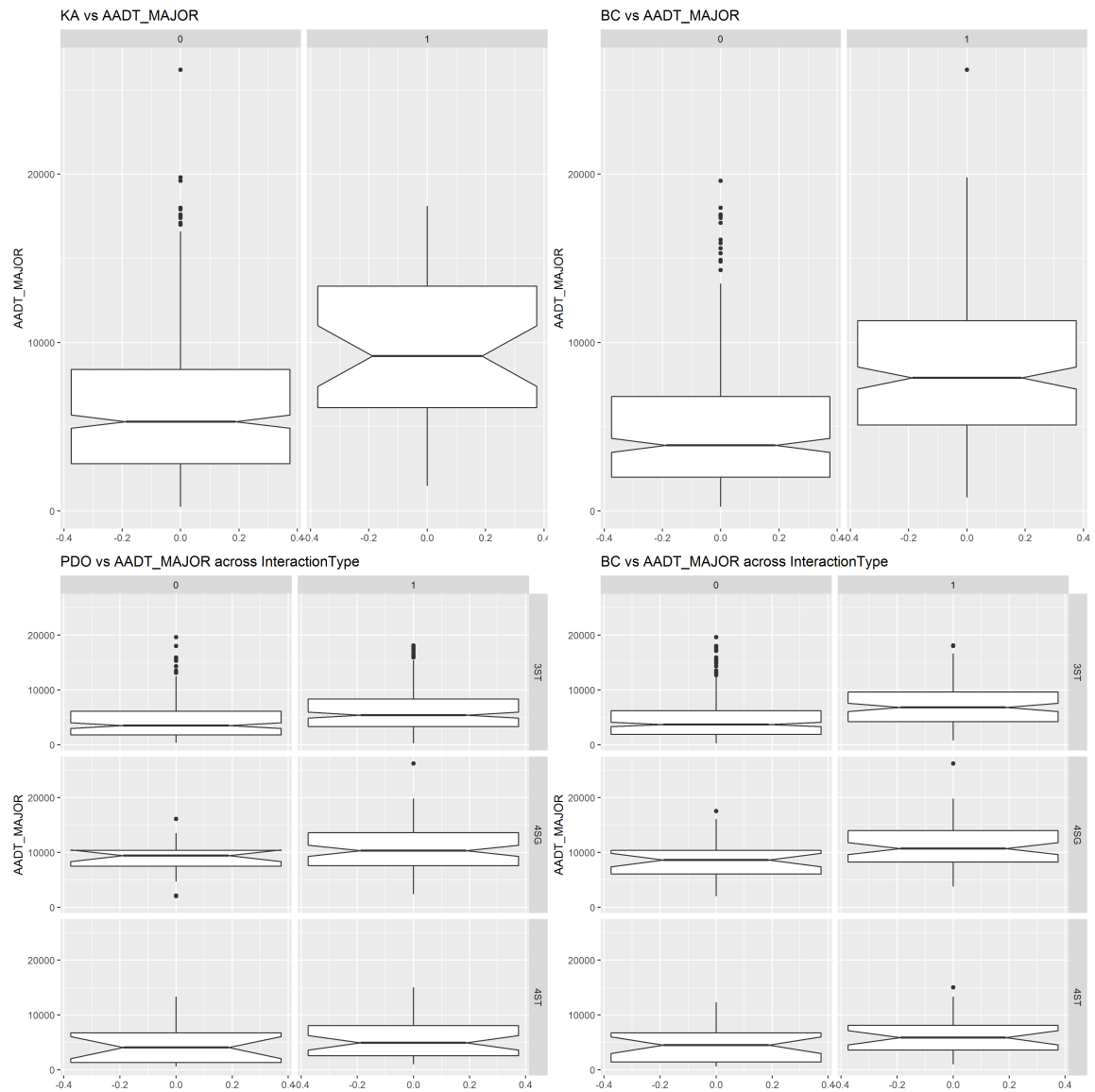


Figure 6.1: Extra plots for AADT-Major

Extra support plots for AADT-Minor:

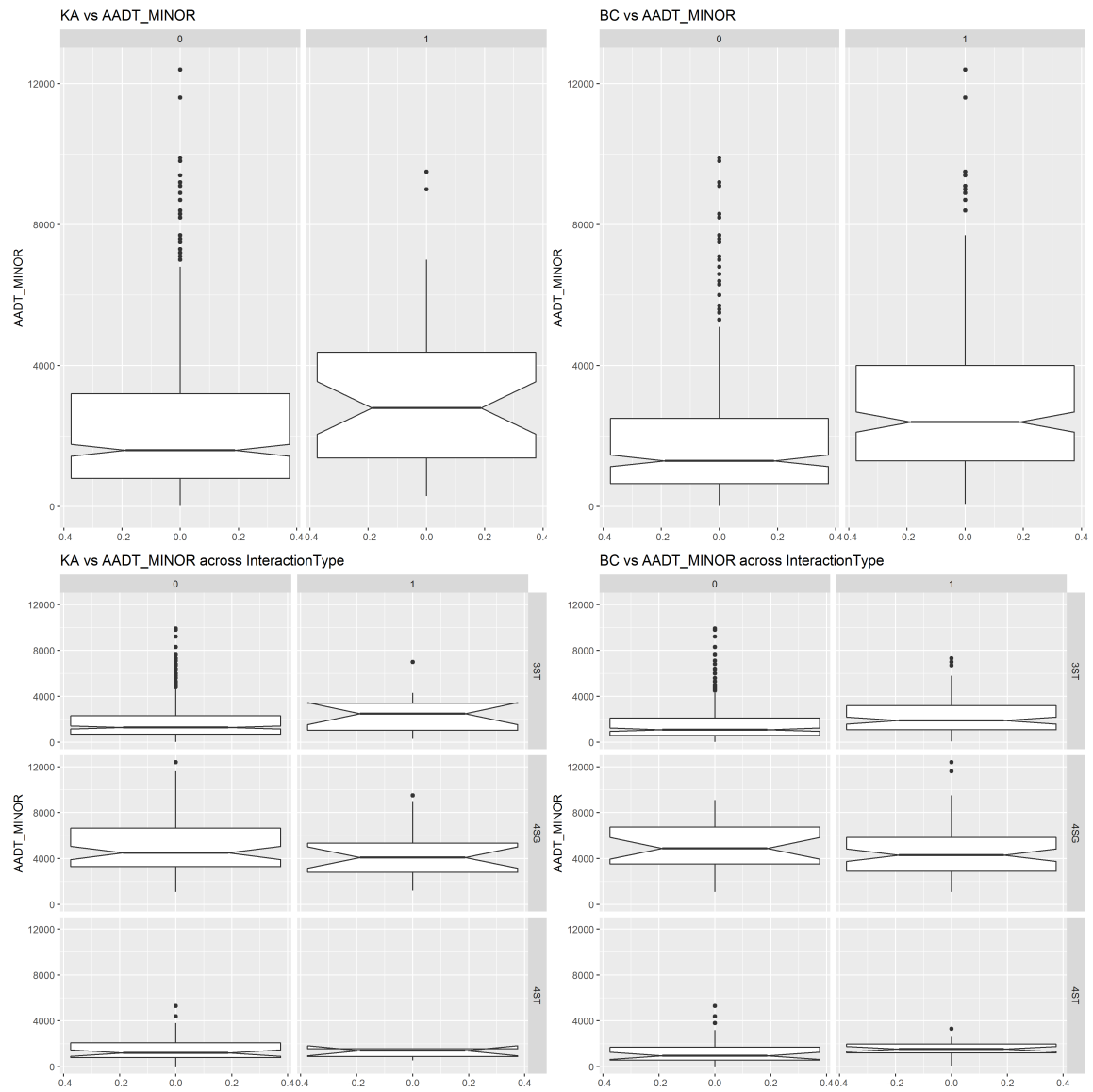


Figure 6.2: Extra plots for AADT-Minor

Extra support plots for Skew Angle:

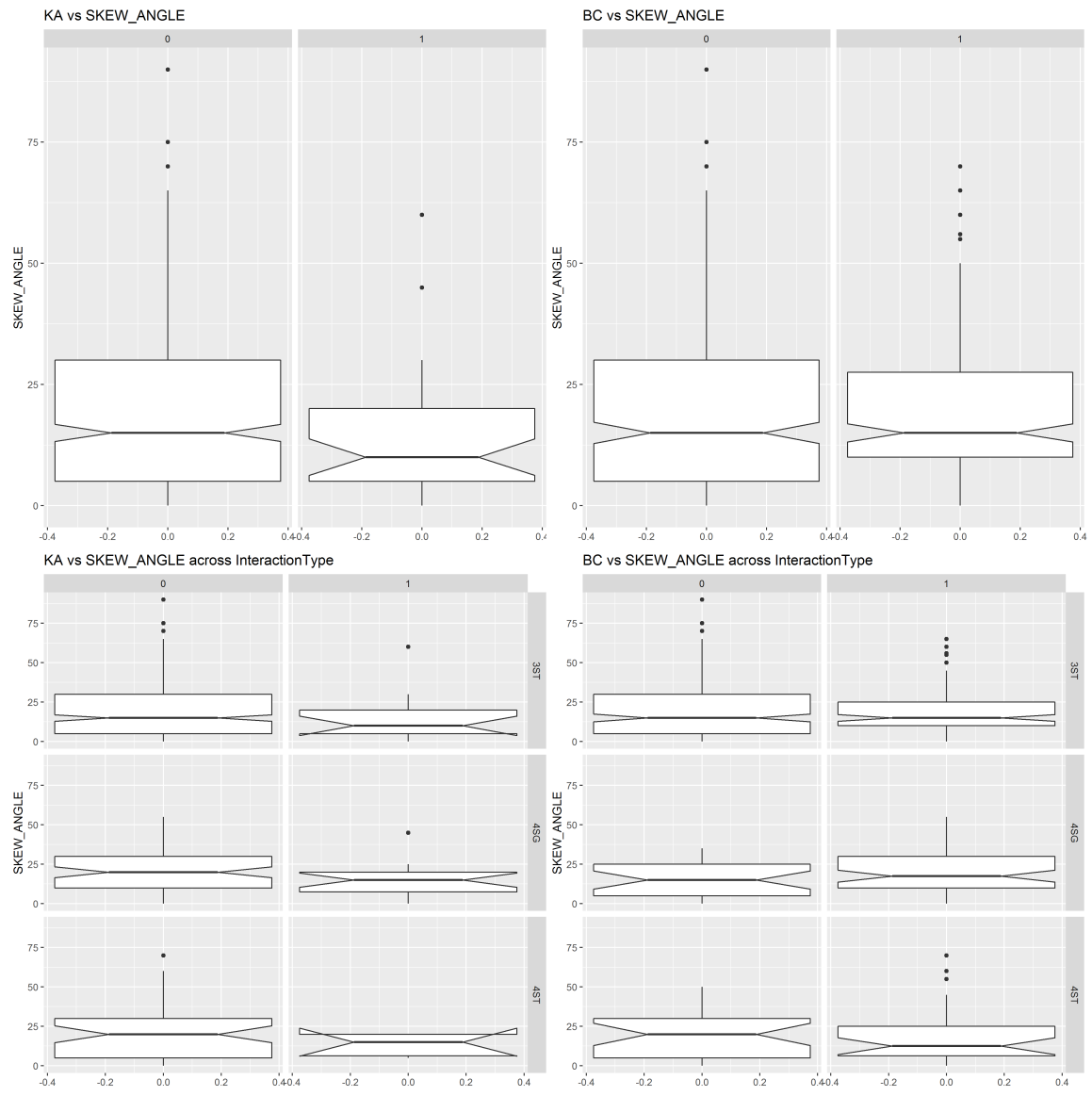


Figure 6.3: Extra plots for Skew-Angle

Extra support plots for Lighting:

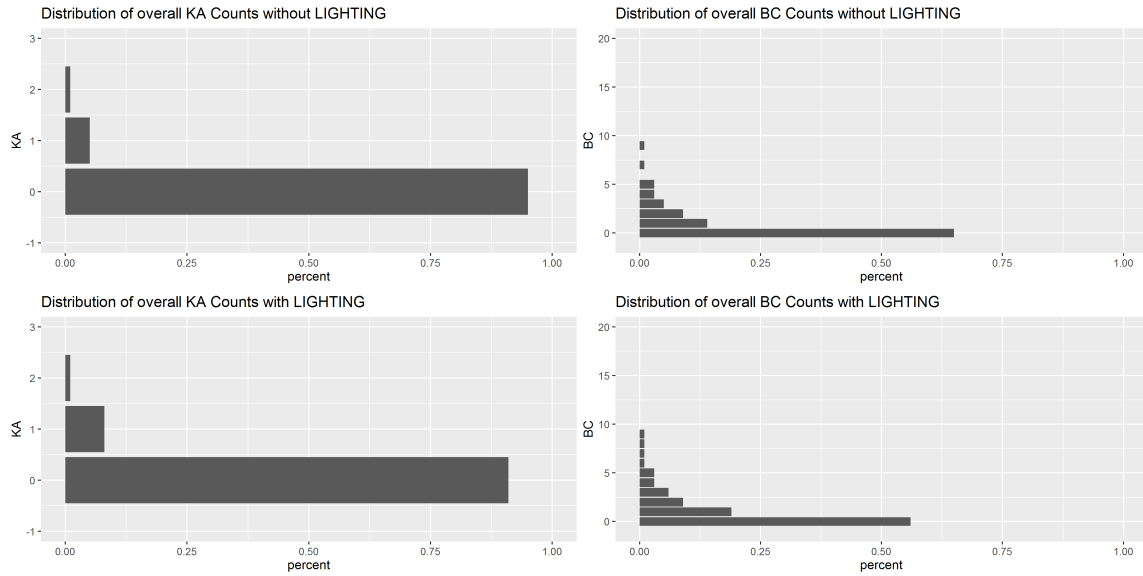


Figure 6.4: Extra plots for Lighting

## 6.2 R codes

## REFERENCES

- [1] LI, BING & ARTEMIOU, ANDREAS & LI, LEXIN(2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*. **3182–3210**.
- [2] XIA, QIANG & XU, WANGLI & ZHU, LIXING(2015). Consistently determining the number of factors in multivariate volatility modeling. *Statistica Sinica*. **1025–1044**.