Metro Interstate Traffic Volume

Chen Lou and Josef Pishek (Group 4)

MA 5790 Predictive Modeling

UCI Machine Learning Repository Dataset of Hourly Minneapolis-St Paul, MN traffic volume for westbound I-94. Includes weather and holiday features from 2012-2018.

Outline

Background

Data Structure

Objective

Initial Observations

Filtering

Summary Statistics

Dummy Variables

Resampling Methods

References

Automatic Traffic Recorders (ATR)

One of several methods for collecting data on traffic volume.

Permanent installations with varying levels of technology to continuously monitor traffic volume, additional types of data depending upon their equipment and sensors.

This project's data is from one of 70+ active devices in Minnesota-- 30+ in the seven-county metro area and 35+ in greater Minnesota.

Find more information about Traffic Forecasting and Analysis from the Minnesota Department of Transportation (MNDoT).

Data Structure

holiday	Categorical US National holidays plus regional holiday, Minnesota State Fair
temp	Numeric Average temp in kelvin
rain_1h	Numeric Amount in mm of rain that occurred in the hour
snow_1h	Numeric Amount in mm of snow that occurred in the hour
clouds_all	Numeric Percentage of cloud cover
weather_main	Categorical Short textual description of the current weather
weather_description	Categorical Longer textual description of the current weather
date_time	DateTime Hour of the data collected in local CST time
*traffic_volume	Numeric Hourly I-94 ATR 301 reported westbound traffic volume

- Categorical variables:
  - holiday (12 levels)
  - weather_main (11 levels)
  - weather_description (38 levels)
- Numerical variables:
  - temp
  - rain_1h
  - snow_1h
  - clouds_all
  - traffic_volume
- Date and Time variables:
  - date_time
- Sample size: 9 variables with 48204 observations
- *Response variable: traffic_volume
- Source notes: MNDoT ATR station 301, roughly midway between Minneapolis and St Paul, MN. Weather data from OpenWeatherMap.

Objective

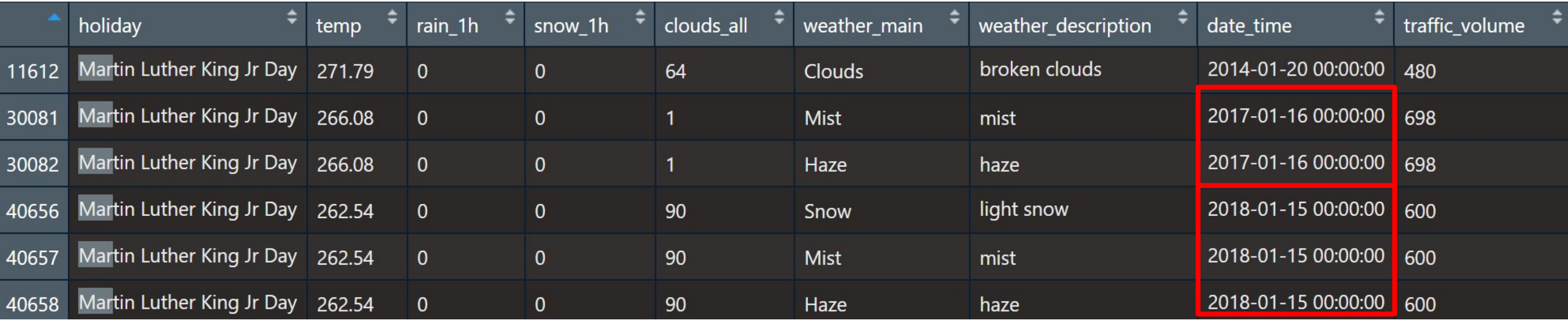
Predict the response variable “traffic_volume” from a collection of numerical and categorical predictors by:

- **Preprocessing Data**
- Fitting and Evaluating a Model

Initial Observations

There are no missing data, but there is repeated data.

Date and Time need to be converted to categorical attributes: year, month, day of week, and hour.



	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume
11612	Martin Luther King Jr Day	271.79	0	0	64	Clouds	broken clouds	2014-01-20 00:00:00	480
30081	Martin Luther King Jr Day	266.08	0	0	1	Mist	mist	2017-01-16 00:00:00	698
30082	Martin Luther King Jr Day	266.08	0	0	1	Haze	haze	2017-01-16 00:00:00	698
40656	Martin Luther King Jr Day	262.54	0	0	90	Snow	light snow	2018-01-15 00:00:00	600
40657	Martin Luther King Jr Day	262.54	0	0	90	Mist	mist	2018-01-15 00:00:00	600
40658	Martin Luther King Jr Day	262.54	0	0	90	Haze	haze	2018-01-15 00:00:00	600

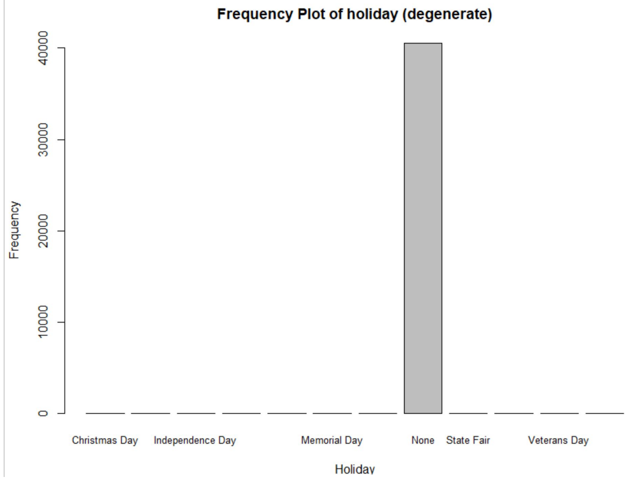
Filtering Duplicate Observations

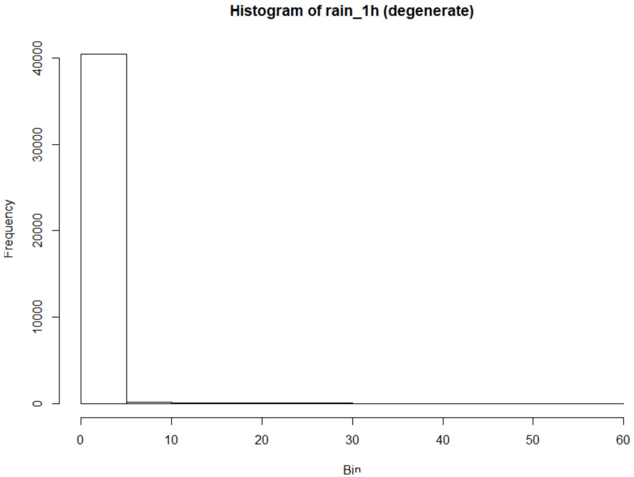
- There are 40,575 unique date_time values of the 48,204 observations
  - Taking the first observation of any duplicated date_time
- There are 35,130 non-duplicated date_time values of the 48,204 observations
  - Removing any observations that have repeated date_time

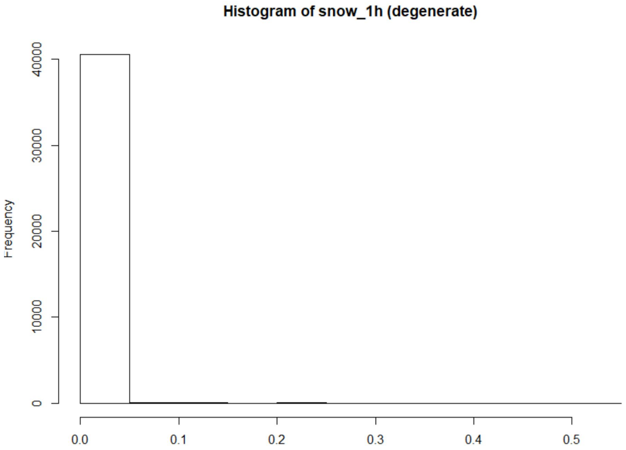
Filtering Near Zero Variance Predictors

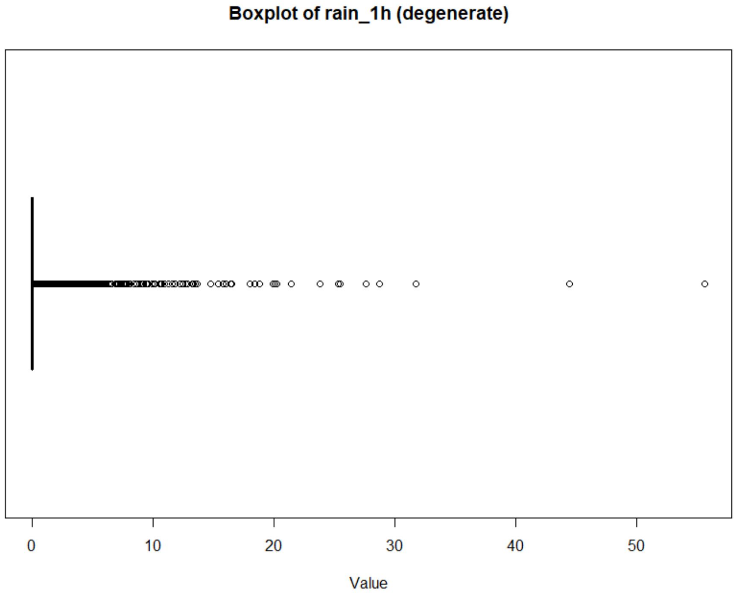
From initial screening, the following predictors exhibited signs of degenerate distributions

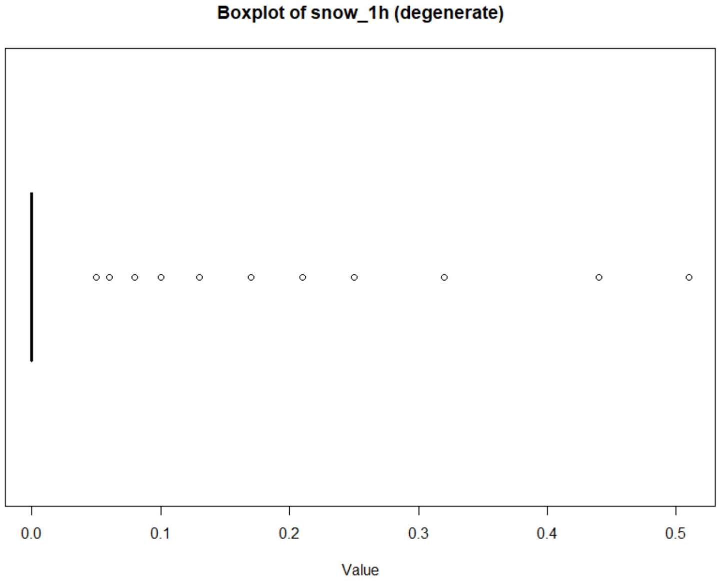
- holiday
- rain_1h
- snow_1h


Frequency Plot of holiday (degenerate)
Frequency
0
10000
20000
30000
40000
Christmas Day
Independence Day
Memorial Day
None
State Fair
Veterans Day
Holiday

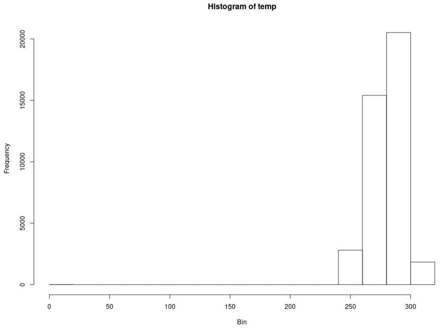

Histogram of rain_1h (degenerate)
Frequency
0
10000
20000
30000
40000
0
10
20
30
40
50
60
Bin

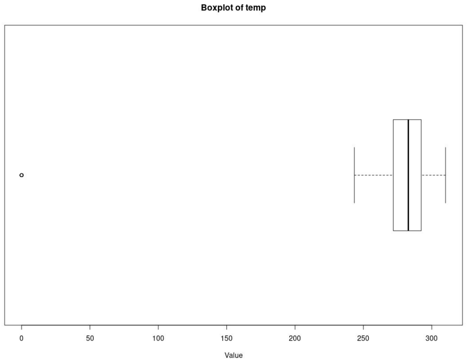

Histogram of snow_1h (degenerate)
Frequency
0
10000
20000
30000
40000
0.0
0.1
0.2
0.3
0.4
0.5


Boxplot of rain_1h (degenerate)
0
10
20
30
40
50
Value

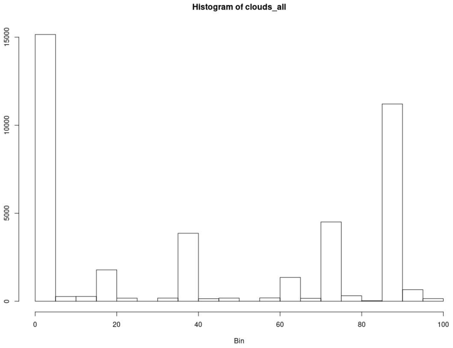

Boxplot of snow_1h (degenerate)
0.0
0.1
0.2
0.3
0.4
0.5
Value

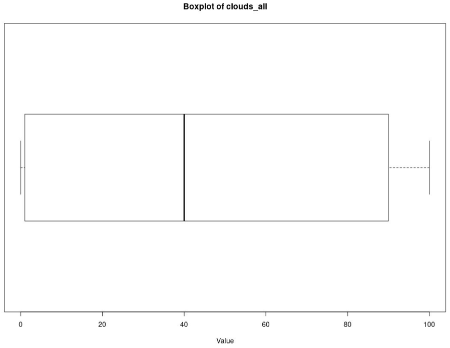
Summary Statistics for Numerical Variables


Histogram of temp

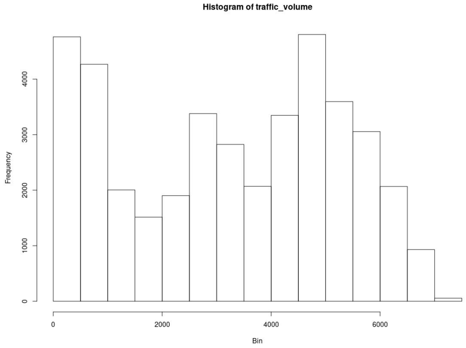

Boxplot of temp

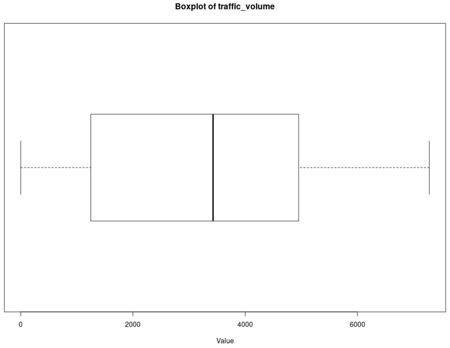
temp


Histogram of clouds_all


Boxplot of clouds_all

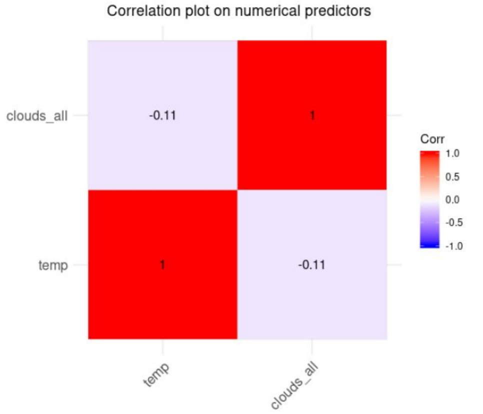
clouds_all

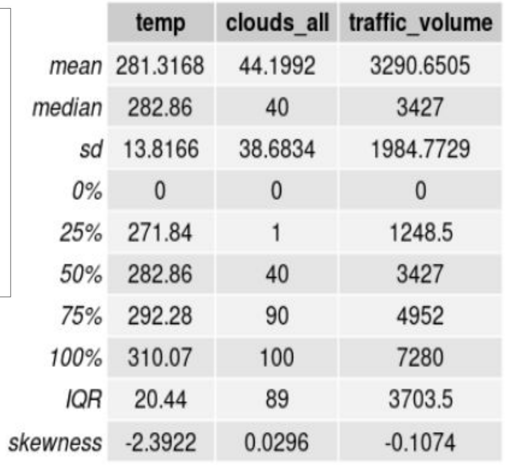

Histogram of traffic_volume


Boxplot of traffic_volume

traffic_volume

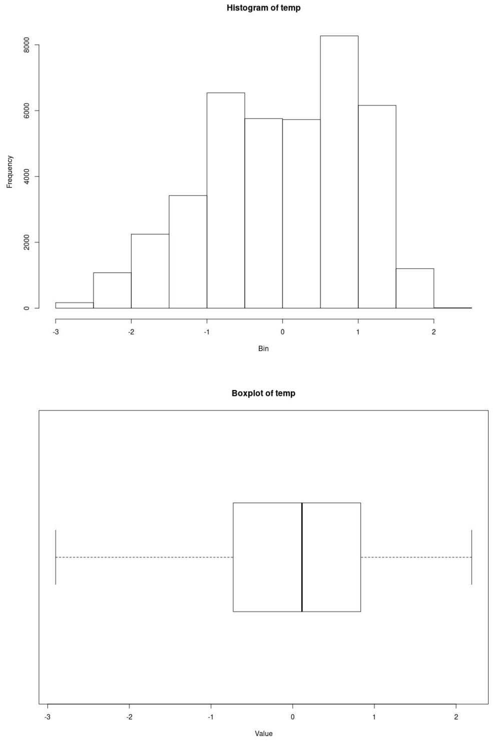
*response


Correlation plot on numerical predictors

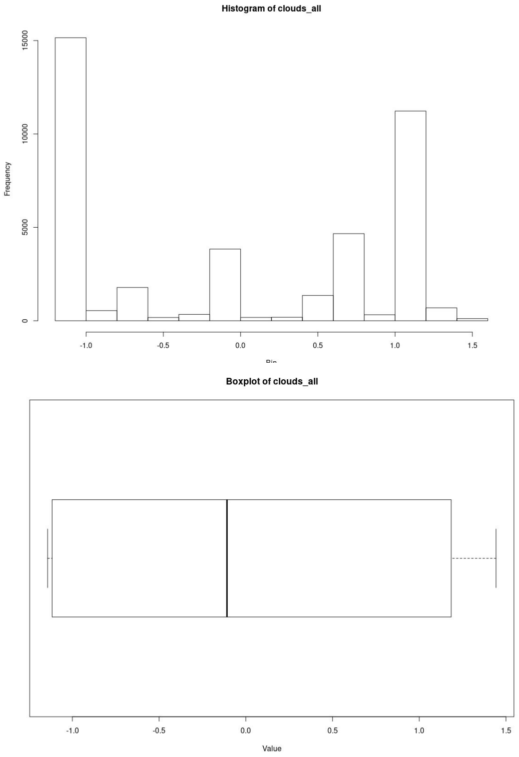


	temp	clouds_all	traffic_volume
mean	281.3168	44.1992	3290.6505
median	282.86	40	3427
sd	13.8166	38.6834	1984.7729
0%	0	0	0
25%	271.84	1	1248.5
50%	282.86	40	3427
75%	292.28	90	4952
100%	310.07	100	7280
IQR	20.44	89	3703.5
skewness	-2.3922	0.0296	-0.1074

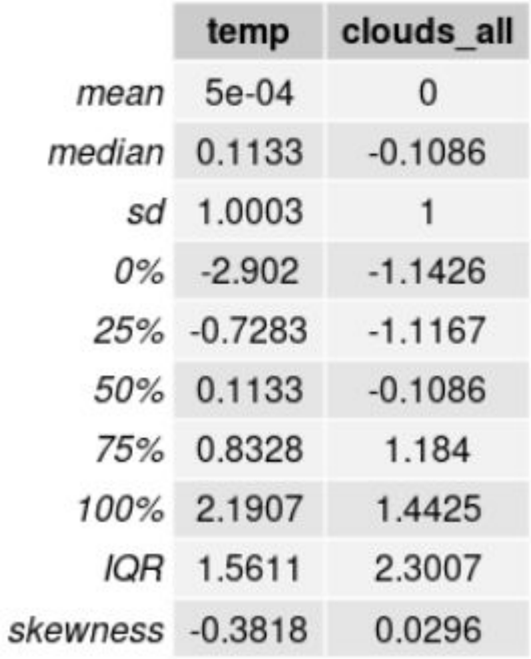
Summary Statistics for Numerical Variables after Center, Scale, KNN Imputation



temp

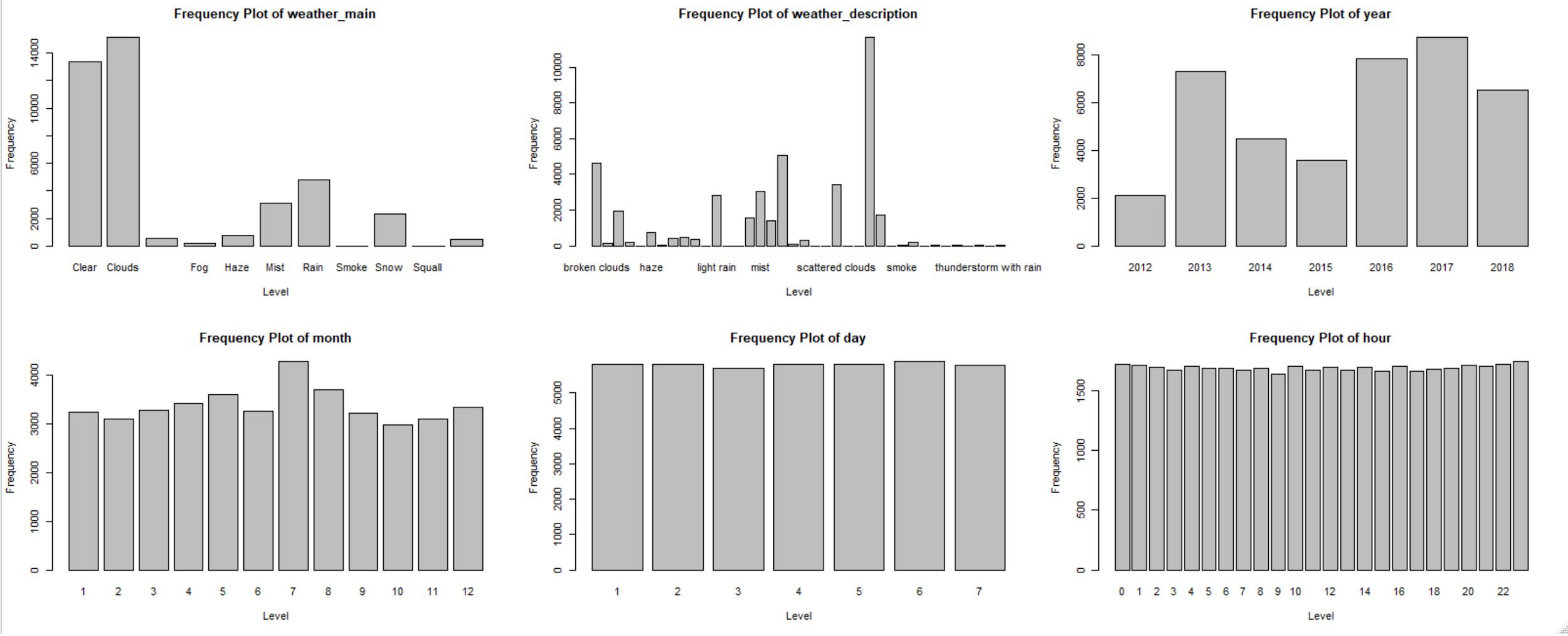


clouds_all

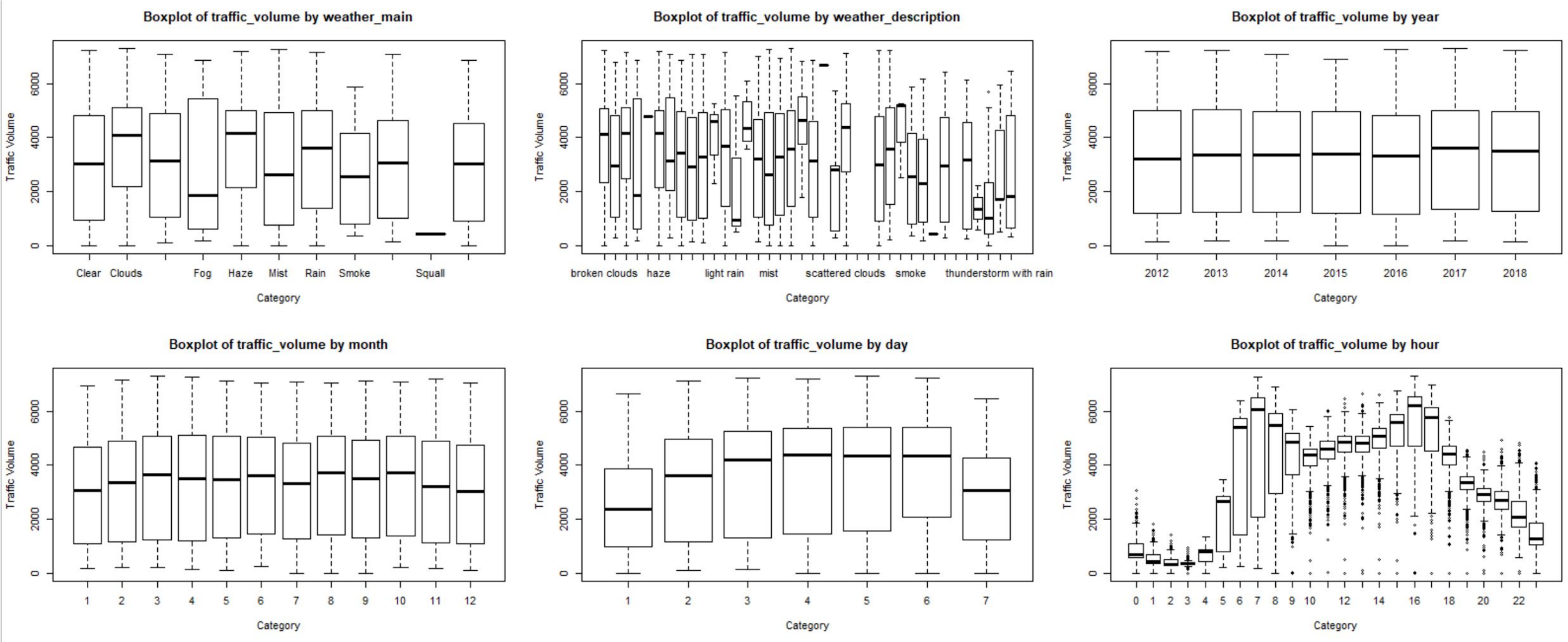


	temp	clouds_all
mean	5e-04	0
median	0.1133	-0.1086
sd	1.0003	1
0%	-2.902	-1.1426
25%	-0.7283	-1.1167
50%	0.1133	-0.1086
75%	0.8328	1.184
100%	2.1907	1.4425
IQR	1.5611	2.3007
skewness	-0.3818	0.0296

Summary Statistics for Categorical Variables



Response by Categorical Predictors

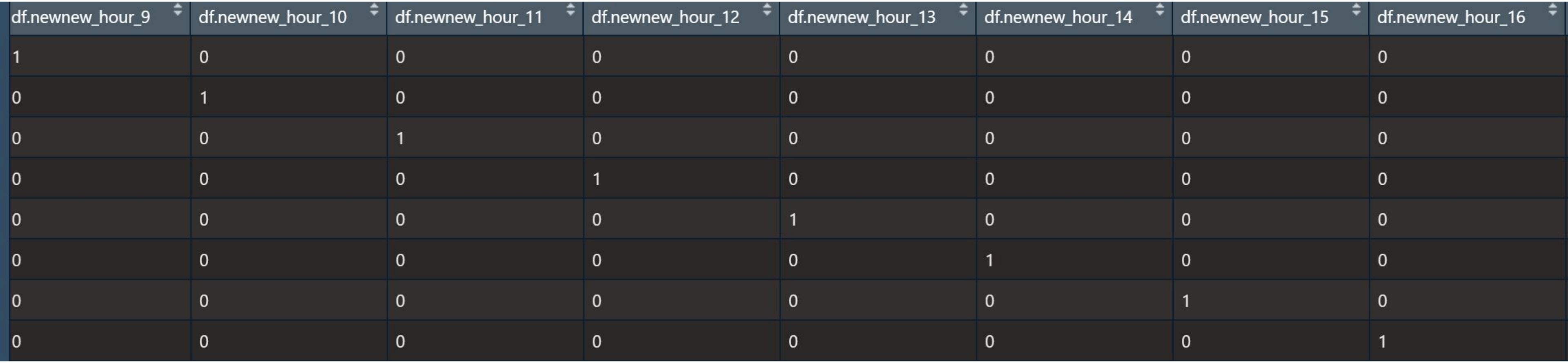


Sig diff? -> ANOVA

Dummy Variables

Created a binary predictor for each level of each categorical predictor.

With weather_main (11), weather_description (35), year (7), month (12), day (7), hour (24), plus 2 numerical predictors (temp and clouds_all) yields 98 predictors.



df.newnew_hour_9	df.newnew_hour_10	df.newnew_hour_11	df.newnew_hour_12	df.newnew_hour_13	df.newnew_hour_14	df.newnew_hour_15	df.newnew_hour_16
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1

Resampling Methods

Generic 80% Train, 20% Test sets from 40,575 observations

10-fold validation: Our dataset is large; low bias; time consuming

- Do not need to repeat, because of large sample size

Bootstrap validation: Relies on random sampling replacement; low variances

References

Hogue, John. (2019). “Metro Interstate Traffic Volume Data Set”. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

Kuhn M, Johnson K (2013). Applied Predictive Modeling. Springer Science + Business Media.

MNDoT (2019). “Collection Methods”. Traffic Forecasting and Analysis. St. Paul, MN.