Data 621: Assignment 3

*Binary Logistic Regression*

*Ritesh Lohiya*

*June 30, 2018*

## *Overview*

**In this homework assignment, you will explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).**

**Your objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation data set using your binary logistic regression model. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:**

**Data Exploration:**

Our training data comprises 466 observations and 13 variables. Below is a brief description of the variables in our data set:

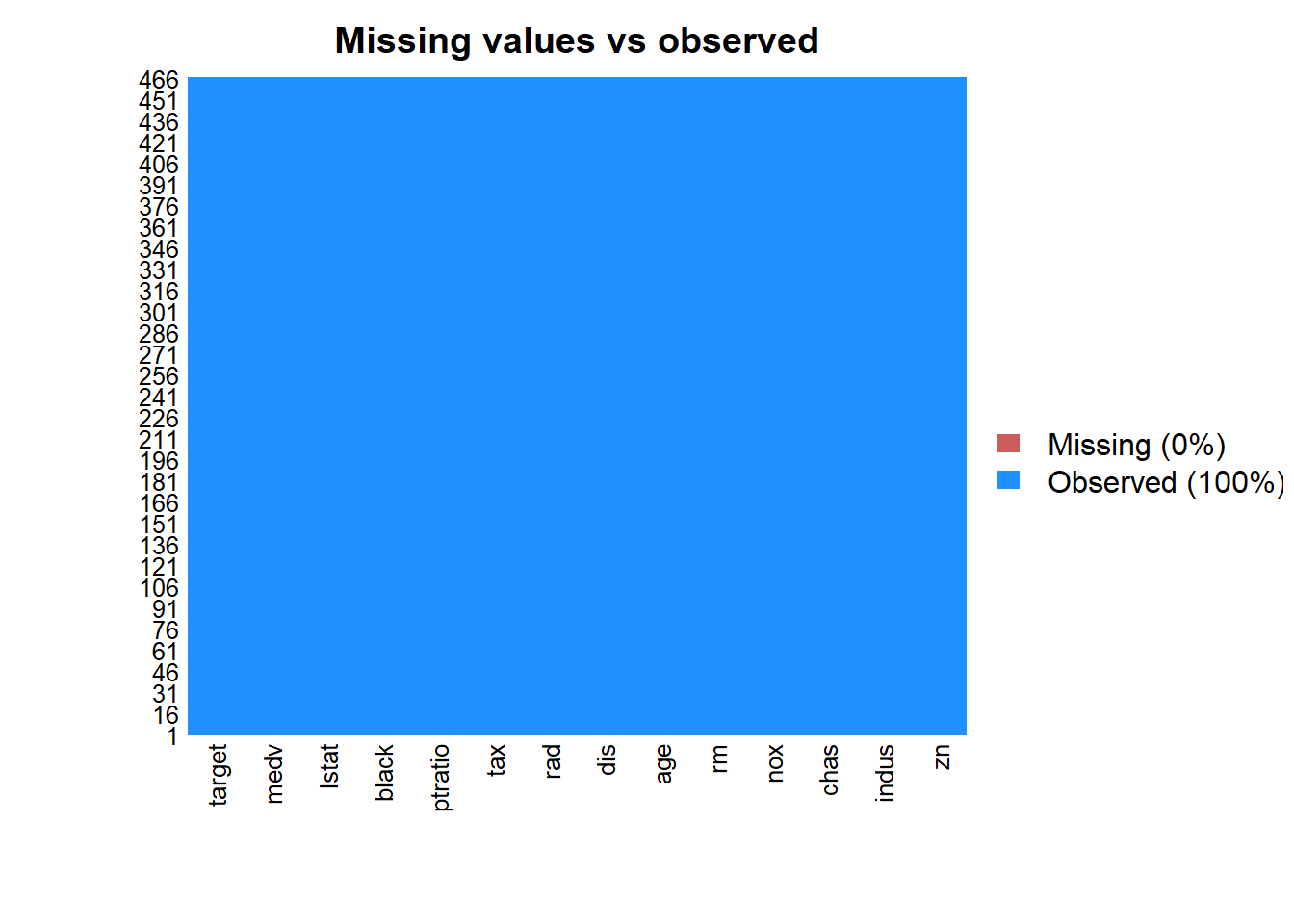
| **Variable Name** | **Description** | **Variable Type** |
| --- | --- | --- |
| zn | proportion of residential land zoned for large lots (over 25000 square feet) | predictor |
| indus | proportion of non-retail business acres per suburb | predictor |
| chas | a dummy var. for whether the suburb borders the Charles River (1) or not (0) | predictor |
| nox | nitrogen oxides concentration (parts per 10 million) | predictor |
| rm | average number of rooms per dwelling | predictor |
| age | proportion of owner-occupied units built prior to 1940 | predictor |
| dis | weighted mean of distances to five Boston employment centers | predictor |
| rad | index of accessibility to radial highways | predictor |
| tax | full-value property-tax rate per $10,000 | predictor |
| ptratio | pupil-teacher ratio by town | predictor |
| black | 1000 (Bk−0.63)2(Bk−0.63)2 where Bk is the proportion of blacks by town | predictor |
| lstat | lower status of the population (percent) | predictor |
| medv | median value of owner-occupied homes in $1000s | response |
| target | whether the crime rate is above the median crime rate (1) or not (0) (response variable) | response |

Target is our binary response variable. For this data exploration, we will be focusing on a binary logistic regression.

**Visual Exploration:**

### Missing Values

We will see the missing values in the dataset. For this i have used Amelia package. According to the graph, the data shows no missing variables.



Lets now dig into our response variables.

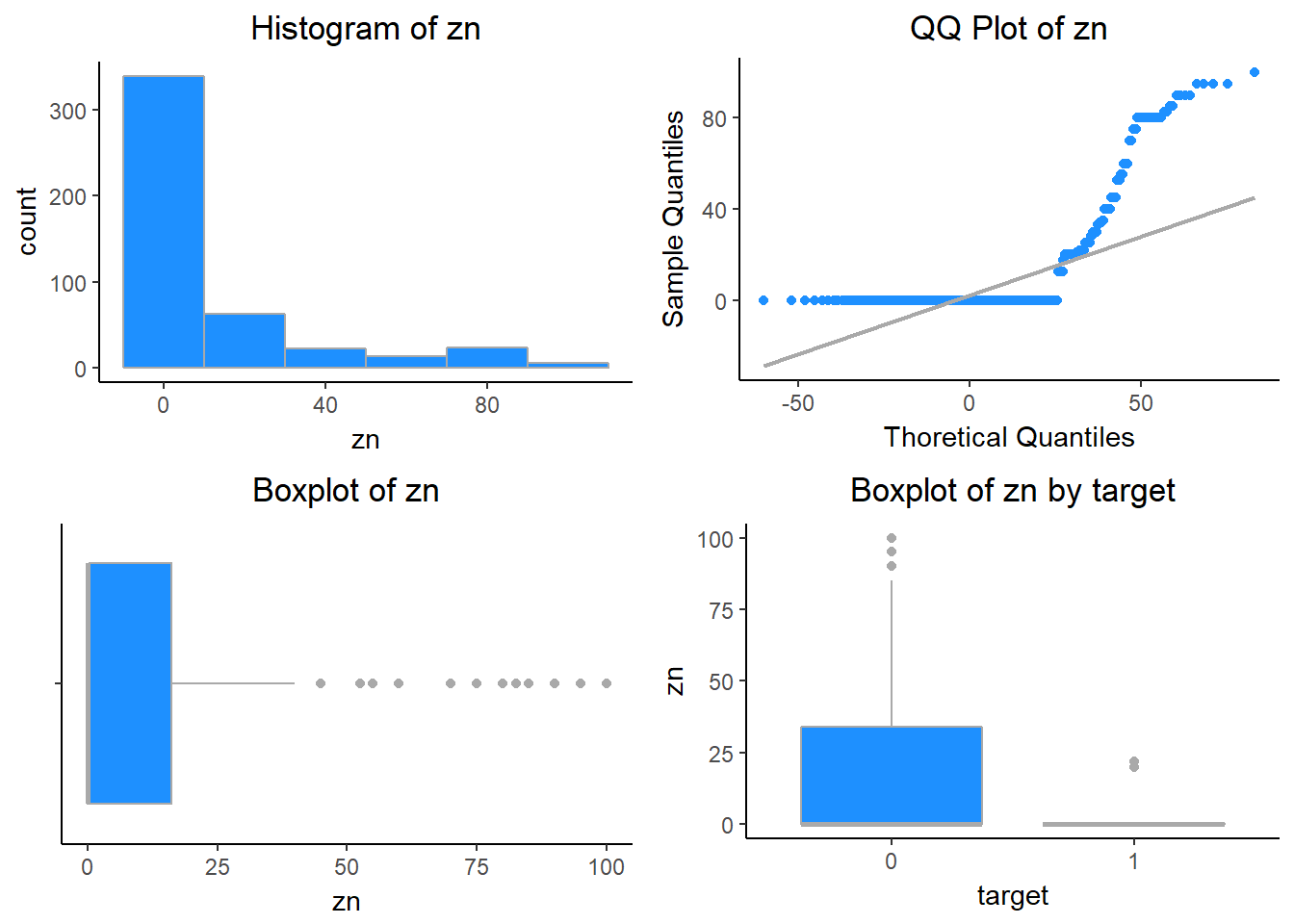
1. Response Variable zn - proportion of residential land zoned for large lots (over 25000 square feet). We can see there are more zeros values for zn and also has positive skewness. Also there appears to be relationship between crime rates and zn.

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.000000 0.000000 0.000000 11.577253 16.250000 100.000000

## SD Skew Kurt

## 23.364651 2.183841 6.842914



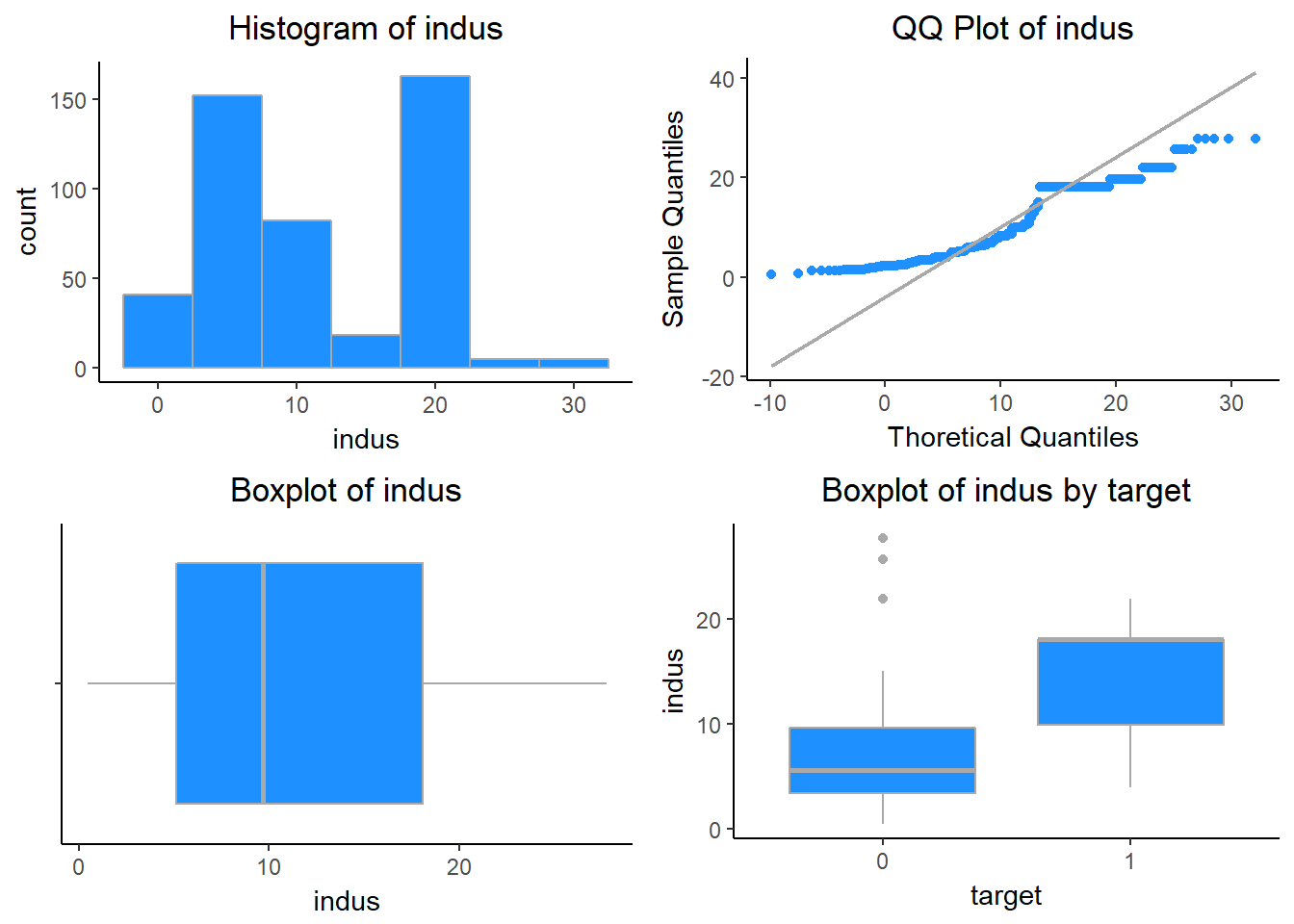
1. Response Variable: indus - proportion of non-retail business acres per suburb. The histogram below indicates a bi-modal quality to the variable’s distribution, with many values clustering in two ranges.

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.4600000 5.1450000 9.6900000 11.1050215 18.1000000 27.7400000

## SD Skew Kurt

## 6.8458549 0.2894763 1.7643510

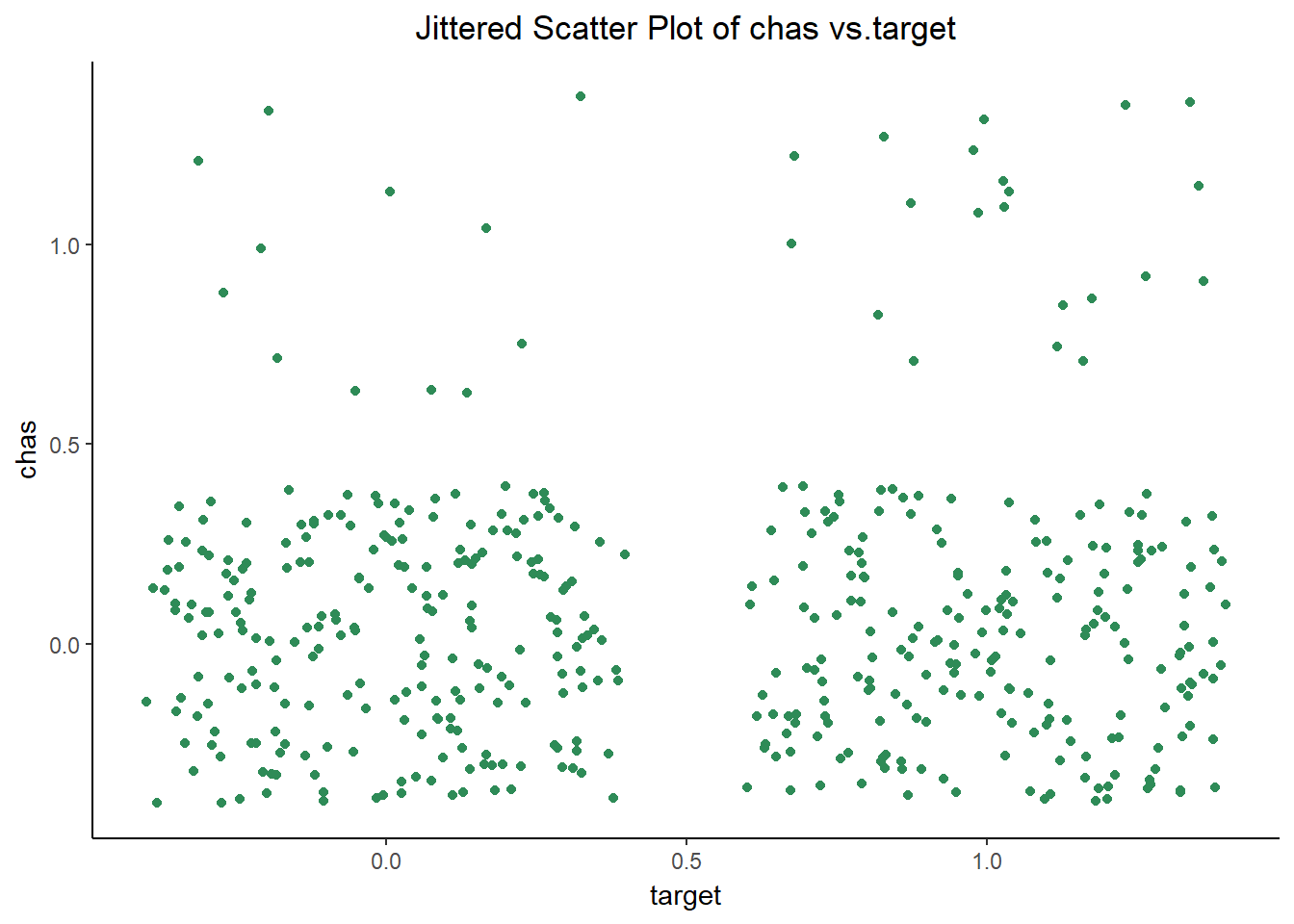


1. Response Variable: chas - a dummy var. for whether the suburb borders the Charles River (1) or not (0). This variable tells us if the neighborhood borders the Charles River (1) or not (0). Close to 7% of the neighborhood borders the Charles River. Of the areas bordering the Charles river 21 are in high crime areas.

##

## 0 1 Sum

## 433 33 466



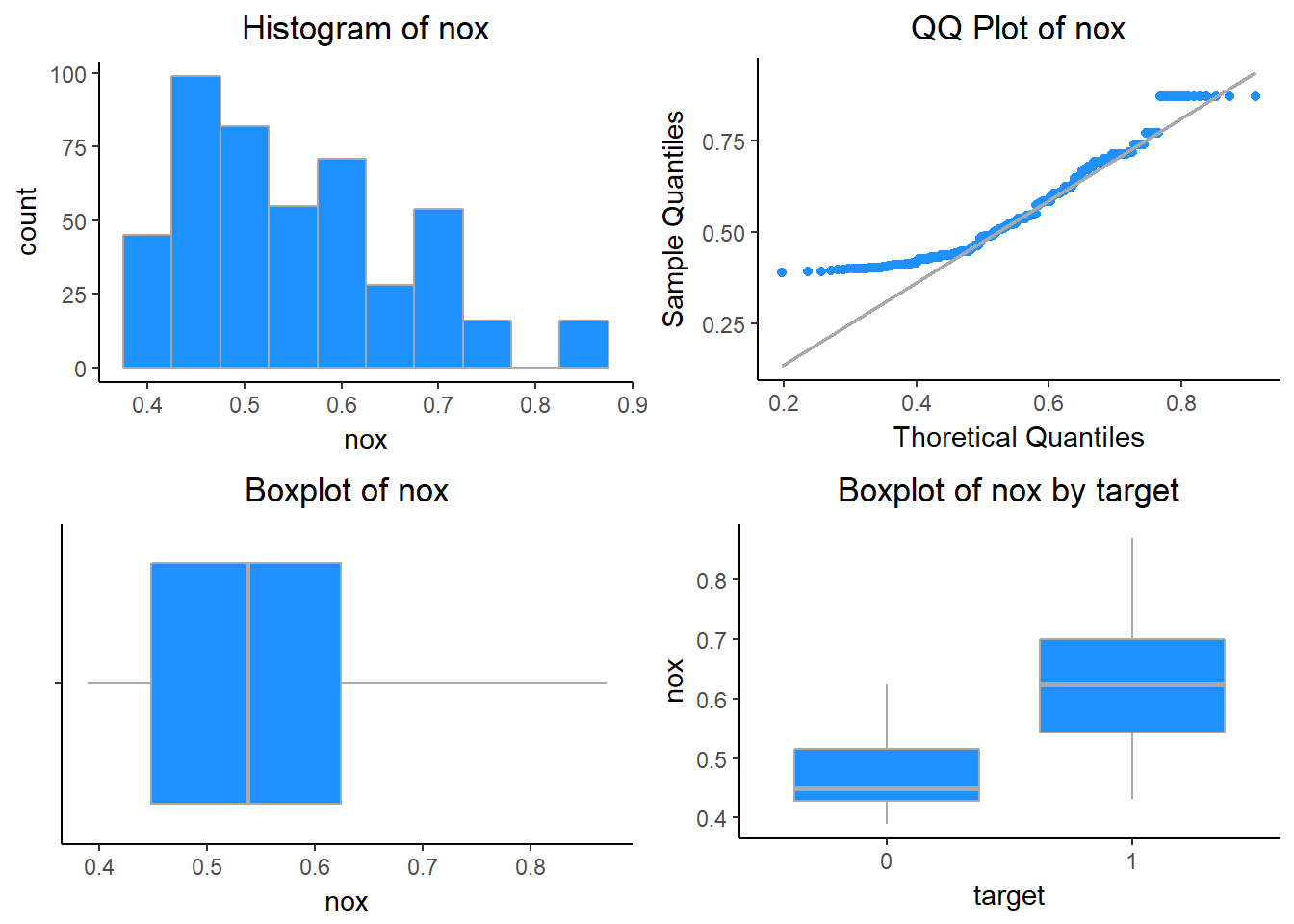
1. Response Variable: nox - nitrogen oxides concentration (parts per 10 million). The variable nox represents the concentration of nitrogen oxide in each Boston area. There is also positive skewness. We also see moderately higher nox variance in high crime areas.

## Min. 1st Qu. Median Mean 3rd Qu. Max. SD

## 0.3890000 0.4480000 0.5380000 0.5543105 0.6240000 0.8710000 0.1166667

## Skew Kurt

## 0.7487369 2.9769895



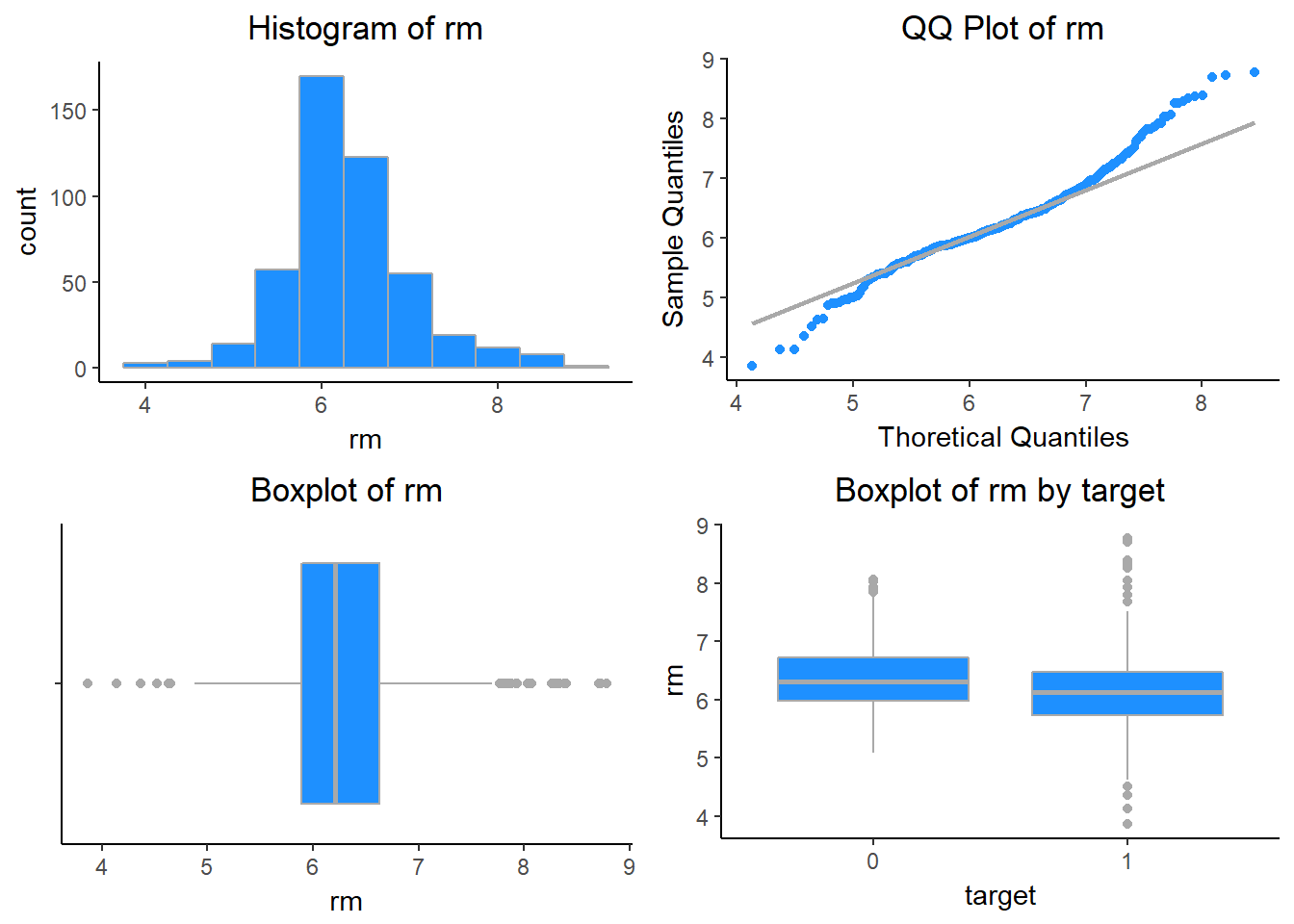
5. Response Variable: rm - average number of rooms per dwelling. The predictor rm is count measure describing the average number of rooms per dwelling. The distridution has heavy tail and has bell curve.

## Min. 1st Qu. Median Mean 3rd Qu. Max. SD

## 3.8630000 5.8872500 6.2100000 6.2906738 6.6297500 8.7800000 0.7048513

## Skew Kurt

## 0.4808673 4.5619962



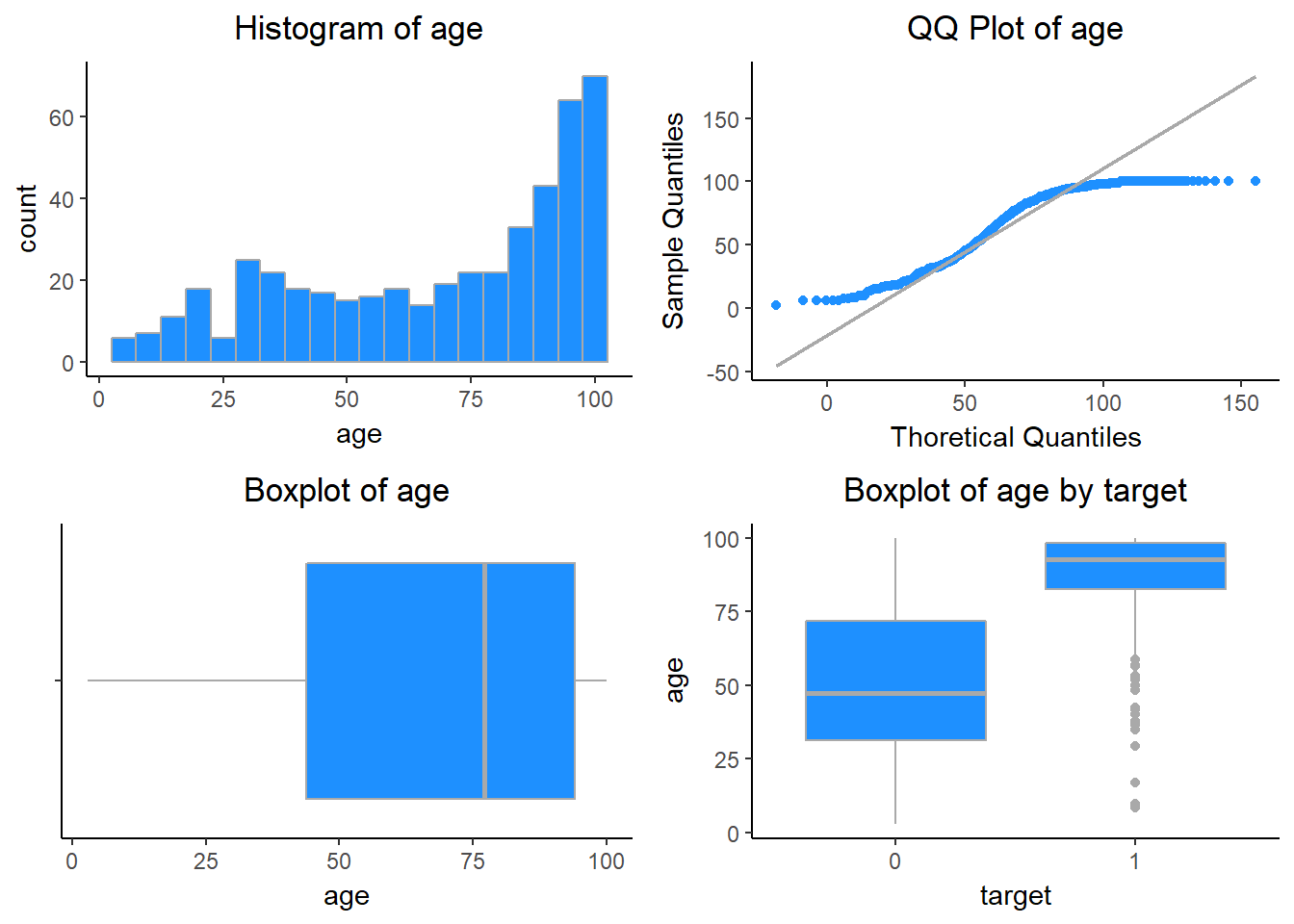
1. Response Variable: age - proportion of owner-occupied units built prior to 1940. The variable age indicates the proportion of owner occupied units built prior to 1940. This variable has high left skewness. Also there is significantly higher mean percentage of older homes in high crime areas.

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 2.9000000 43.8750000 77.1500000 68.3675966 94.1000000 100.0000000

## SD Skew Kurt

## 28.3213784 -0.5795721 1.9986874



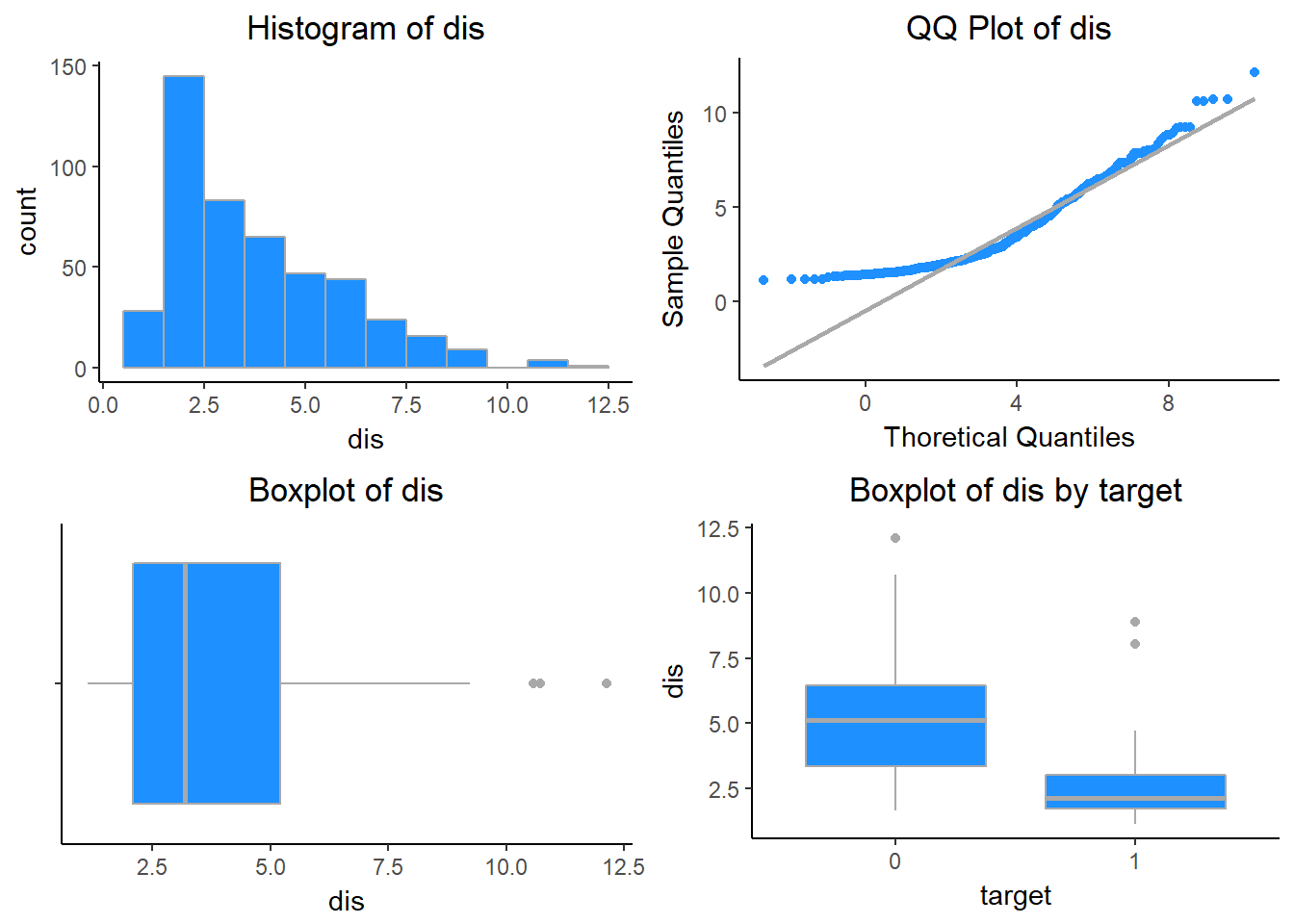
1. Response Variable: dis - weighted mean of distances to five Boston employment centers. The predictor dist describes the average distance to Boston employment centers. The variable is moderately right skewed. Also we can see that low crime areas are associated with higher average distances to employment centers.

## Min. 1st Qu. Median Mean 3rd Qu. Max. SD

## 1.129600 2.101425 3.190950 3.795693 5.214600 12.126500 2.106950

## Skew Kurt

## 1.002117 3.486917



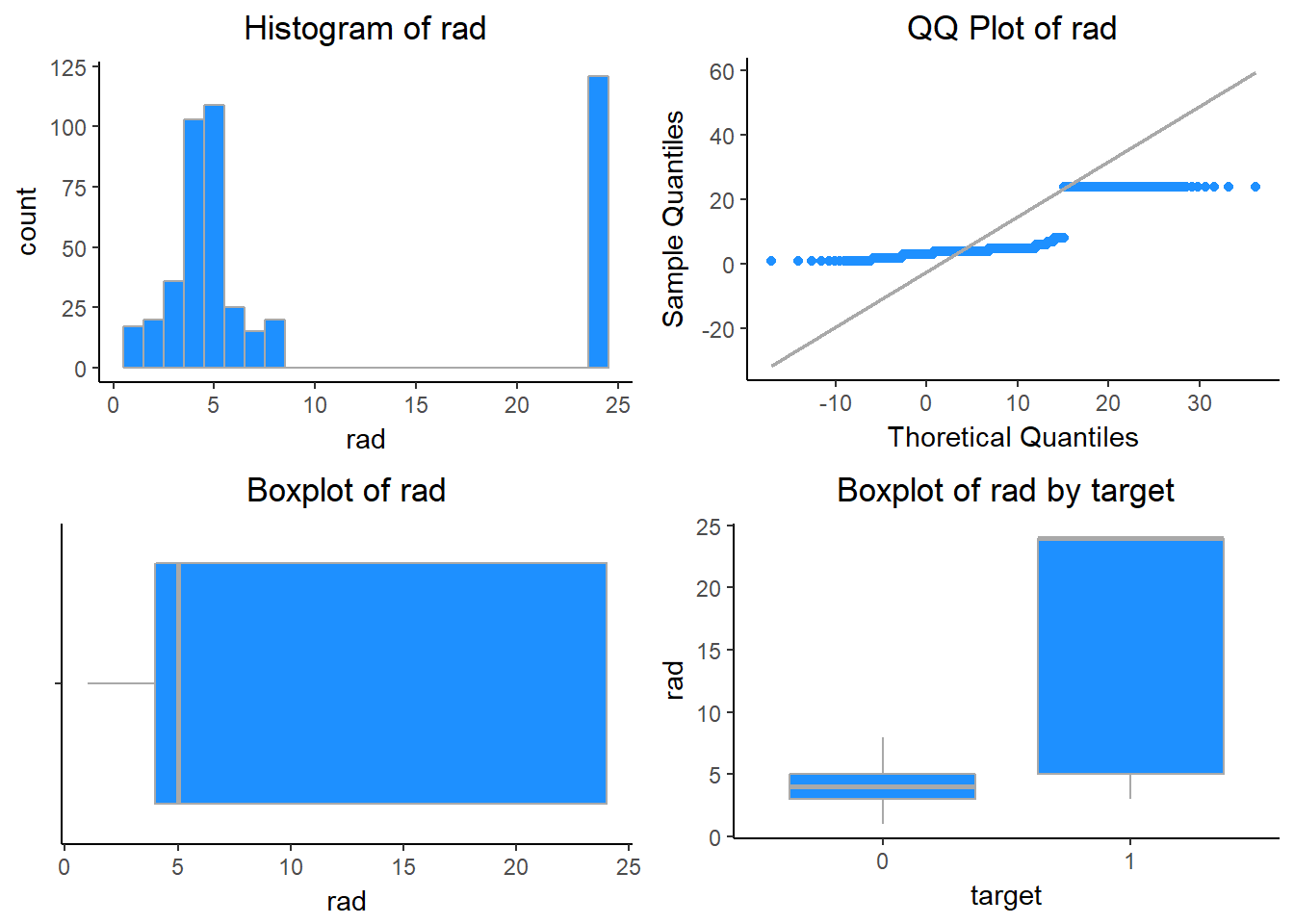
1. Response Variable: rad - index of accessibility to radial highways. The rad variable is an integer-valued index measure indicating an area’s accessibility to radial highways. In the boxplots below, there appears to be a significant positive association between high crime rates and rad value.

## Min. 1st Qu. Median Mean 3rd Qu. Max. SD

## 1.000000 4.000000 5.000000 9.530043 24.000000 24.000000 8.685927

## Skew Kurt

## 1.013539 2.147295



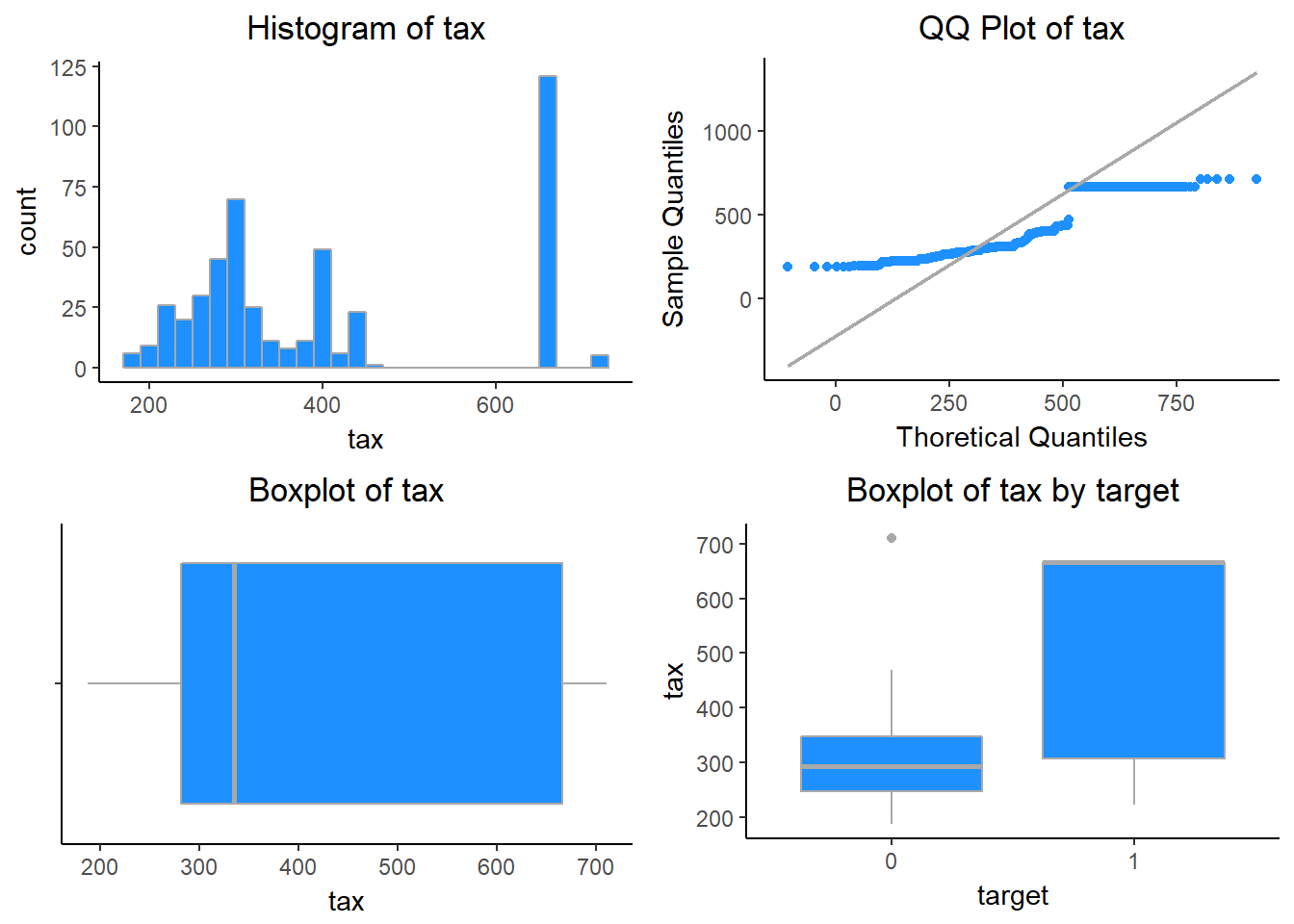
1. Response Variable: tax - full-value property-tax rate per $10,000. The tax variable refers to the the tax rate per $10k of property value. High crime areas also appear to have a strong, positive association with the tax value. This variable is densely distributed around two of the following approximate values: 300 and 700.

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 187.0000000 281.0000000 334.5000000 409.5021459 666.0000000 711.0000000

## SD Skew Kurt

## 167.9000887 0.6614416 1.8599284



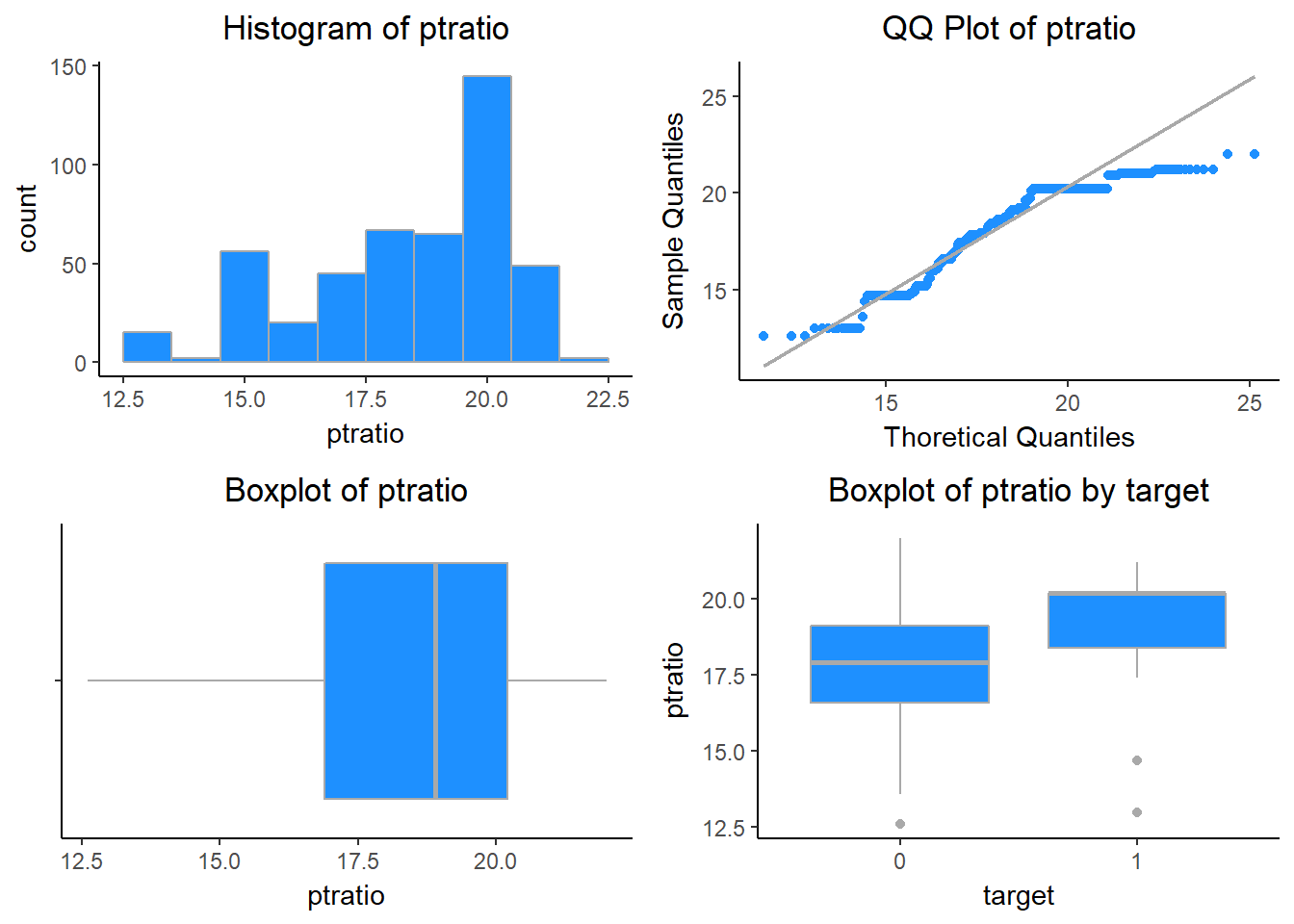
1. Response Variable: ptratio - ptratio: pupil-teacher ratio by town. The predictor ptratio indicates the average school, pupil-to-student ratio, and has a left skewed distribution. We can see a positive relationship between ptratio and high crime.

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 12.6000000 16.9000000 18.9000000 18.3984979 20.2000000 22.0000000

## SD Skew Kurt

## 2.1968447 -0.7567025 2.6108306



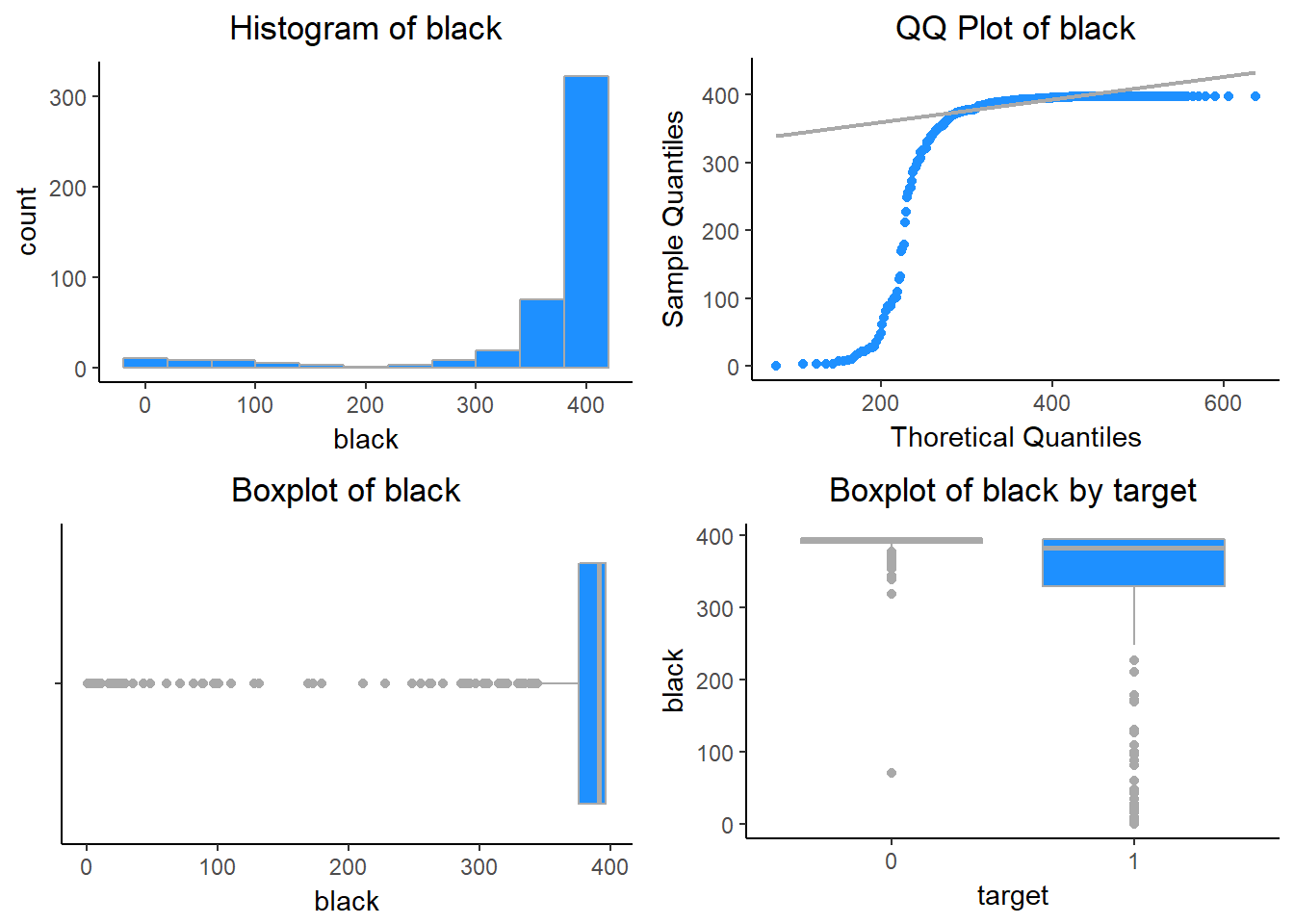
1. Response Variable: black - 1000 (Bk−0.63)2(Bk−0.63)2 where Bk is the proportion of blacks by town. This variable is heavily left skewed.

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 0.320000 375.607500 391.340000 357.120150 396.237500 396.900000

## SD Skew Kurt

## 91.321130 -2.925723 10.386460



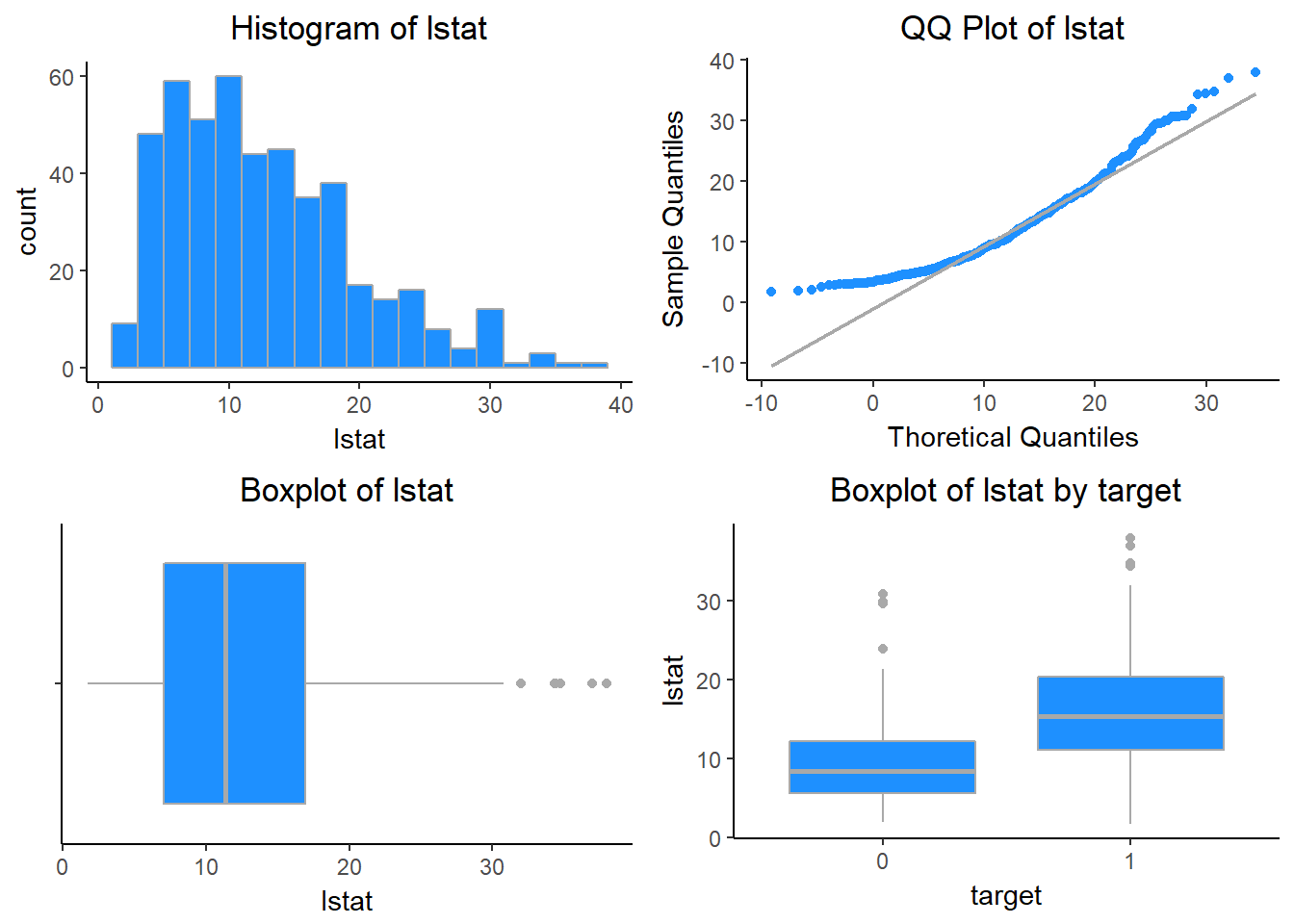
1. Response Variable: lstat - lower status of the population (percent). The variable lstat indicates the proportion of the population deemed to be of lower status. lstat is right skewed. High crime areas tend to have be associated with larger lstat values.

## Min. 1st Qu. Median Mean 3rd Qu. Max.

## 1.7300000 7.0425000 11.3500000 12.6314592 16.9300000 37.9700000

## SD Skew Kurt

## 7.1018907 0.9085092 3.5184532



1. Response Variable: medv - median value of owner-occupied homes in $1000s. The median value of residential homes in a given area. The variable is slightly right skewed, and high values of medv appear to be associated with lower crime rates.

## Min. 1st Qu. Median Mean 3rd Qu. Max. SD

## 5.000000 17.025000 21.200000 22.589270 25.000000 50.000000 9.239681

## Skew Kurt

## 1.080167 4.392615

