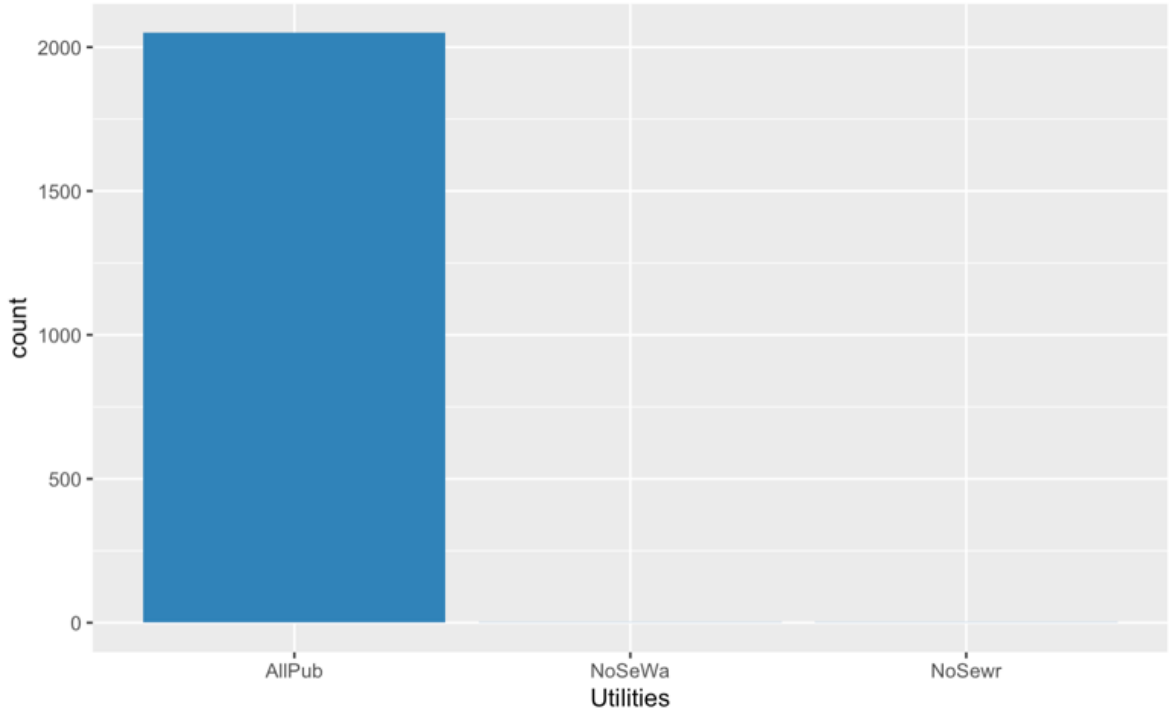
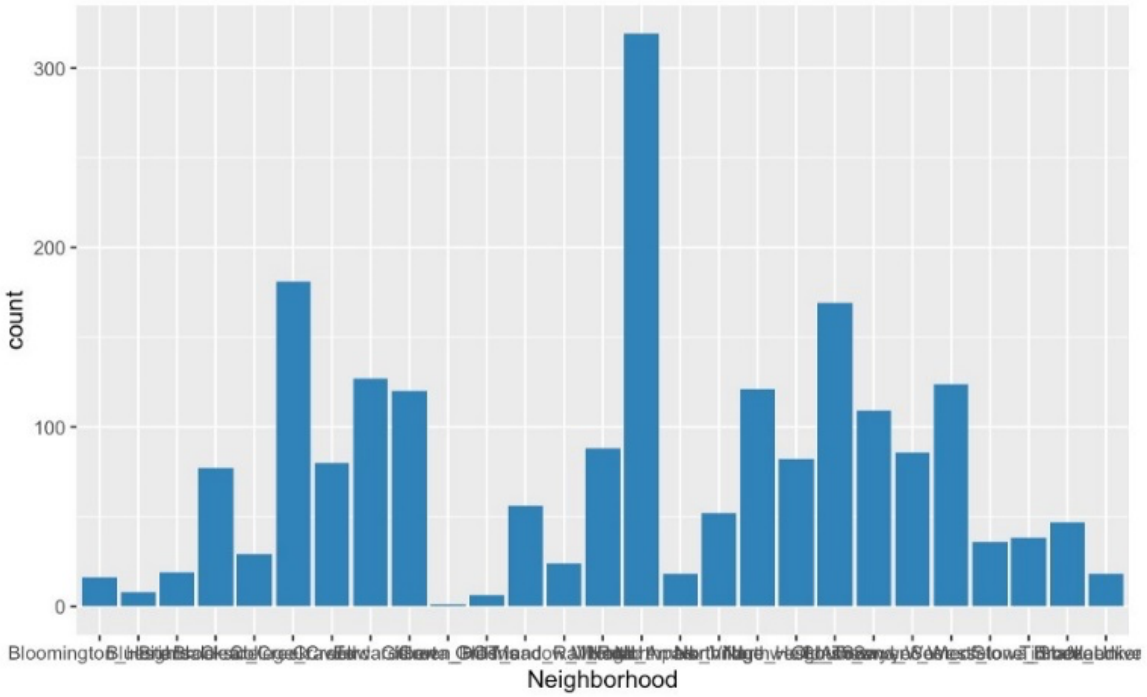
# Appendix

**Imbalance Categorical Variables:** To remove categorical variables where one category may accommodate 95% or more of the total observations, which is not very informative and may cause an imbalance and overfitting. As an example, a distribution of the “Utilities” variable, shows that more than 95% of all the observations were “AllPub” type. Whereas the “Neighborhood” variable had a meaningful distribution that can contribute in the fitting and prediction.



**Winsorization:** Winsorizing a vector means that a predefined quantum of the smallest and/or the largest values are replaced by less extreme values. For example, let's consider the variable "Lot_Frontage". The next plots show the data distribution before and after the winsorization process.