1. Collect and Clean Data

Store the data from the following website in your scripting environment: http://www.utsc.utoronto.ca/~butler/c32/bikes.csv. Then clean the data and return in the following format:

datetime	gender	had_helmet	had_passenger	on_sidewalk
2010-09-24 07:00:00	Male	No	No	No
2010-09-24 07:00:00	Male	No	No	No
2010-09-24 07:00:00	Male	No	No	No
2010-09-24 07:00:00	Male	No	No	No
2010-09-24 07:00:00	Male	No	No	No
2010-09-24 07:15:00	Male	Yes	No	No
2010-09-24 07:15:00	Female	Yes	No	No
2010-09-24 07:15:00	Male	Yes	No	No

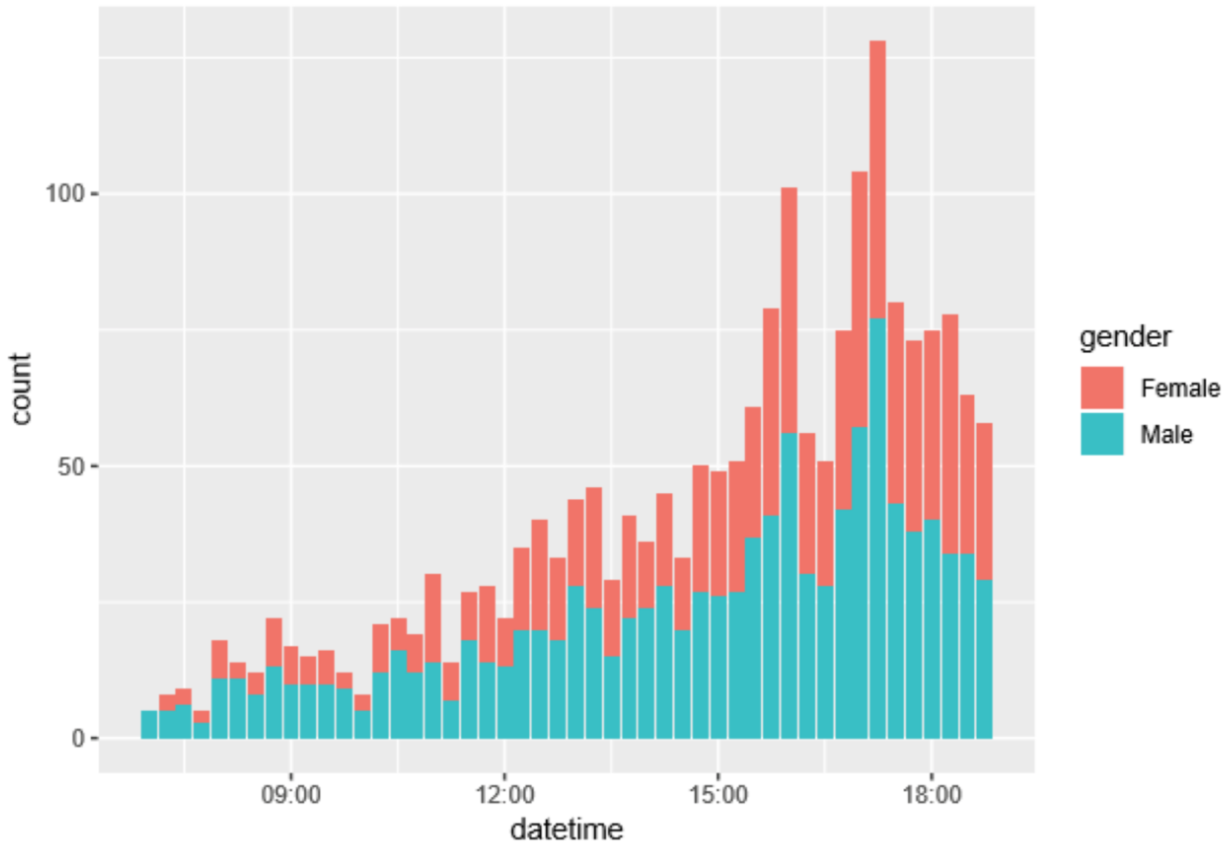
2. Gather Summary Statistics

Report summary statistics of the data similar to the following:

datetime	gender	had_helmet	had_passenger	on_sidewalk
Min. :2010-09-24 07:00:00	Female: 861	No :1007	No :1953	No :1885
1st Qu.:2010-09-24 13:00:00	Male :1097	Yes: 951	Yes: 5	Yes: 73
Median :2010-09-24 15:45:00	NA	NA	NA	NA
Mean :2010-09-24 14:58:05	NA	NA	NA	NA
3rd Qu.:2010-09-24 17:15:00	NA	NA	NA	NA
Max. :2010-09-24 18:45:00	NA	NA	NA	NA

3. Visualization

Display the count of bikers by the time of day, colored by gender, similar to the following:



4. Distribution Fitting

Estimate the probability distribution function for the count of female bikers in the question 3. You need to

1. Hypothesize families of distributions
   - Identify properties of observed data
     - Discrete vs. continuous
     - Bounded, unbounded, non-negative
   - Histogram (plot histogram to visualize the shape of the distribution and choose some candidate distribution functions)
   - Summary statistics (see details below)
2. Estimate the parameters of hypothesized distribution in step 1 (use one of the following two techniques)
   - Maximum likelihood estimators
   - Method of moments
3. Determine the representativeness of each fitted distribution (use one of the following two techniques)
   - Heuristic techniques (frequency comparison, density/histogram plot, cumulative frequency comparison, Q-Q and P-P plot)
   - Goodness-of-fit tests (Chi-square test, K-S test, or Anderson-Darling test)

---Summary statistics---

Population Parameter	*Estimate Statistics*	*Function*	*Distribution*
Min, Max	$x_{(1)}, x_{(n)}$	measure range	C, D
Mean (μ)	$\bar{x}(n) = \dfrac{\sum x_i}{n}$	measure central tendency	C, D
Median ($\tilde{\mu}$)	$\tilde{x}(n) = \begin{cases} x_{(n+1)/2} & n\ \ odd \\ \dfrac{(x_{(n/2)} + x_{((n/2)+1)})}{2} & n\ \ even \end{cases}$	measure central tendency	C, D
Variance (σ^2)	$S^2(n) = \dfrac{\sum (x_i - \bar{x})^2}{n}$	measure variability	C, D
Coefficient of Variation $CV = \sigma / \mu$	$\hat{CV}(n) = \dfrac{S(n)}{\bar{x}(n)}$	alternative measure of variability	C
Lexis ratio $\tau = \sigma^2 / \mu$	$\hat{\tau}(n) = \dfrac{S^2(n)}{\bar{x}(n)}$	measure of variability	D
Skewness $\upsilon = \dfrac{E((x-\mu)^3)}{(\sigma^2)^{3/2}}$	$\hat{\upsilon} = \dfrac{\sum (x_i - \bar{x}(n))^3 / n}{(S^2(n))^{3/2}}$	measure of symmetry	C, D

- If mean = median, symmetric distribution
    - Normal, uniform, triangular distribution
- If CV = 1
    - Exponential distribution
- For Gamma, Weibull and Lognormal distributions

Distribution	**CV**
Gamma	$1/\sqrt{\alpha}$
Weibull	$\dfrac{\left\{2\,\alpha\,\Gamma(2/\alpha) - \left[\Gamma(1/\alpha)\right]^2\right\}^{1/2}}{\Gamma(1/\alpha)}$
Lognormal	$\left(e^{\sigma^2} - 1\right)^{1/2}$

- If CV < 1 and data is skewed to right
    - Gamma or Weibull distribution (with αα > 1)
- If CV > 1 and data is skewed to right
    - Lognormal distribution

- Skewness = 0
  - Symmetric distribution (Normal, uniform, triangular)
- Skewness < 0
  - Data is skewed to left (Triangular or beta)
- Skewness > 0
  - Data is skewed to right (Exponential, lognormal, or gamma and Weibull with $\alpha>1$α>1)