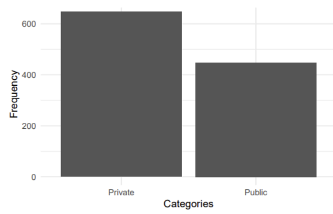
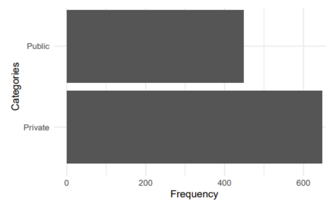
Data Visualization

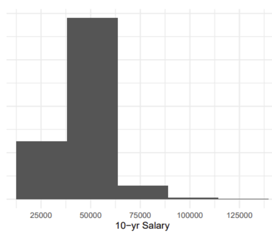
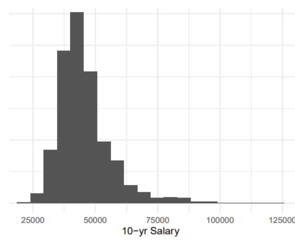
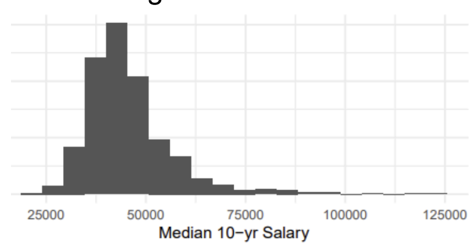
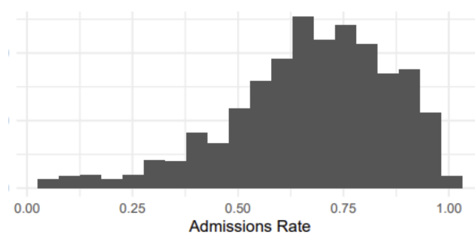
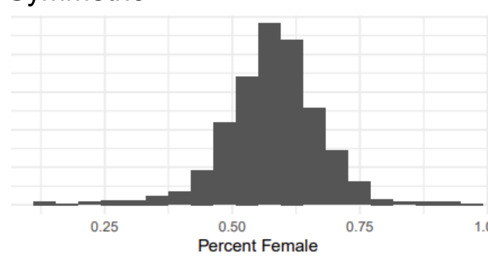
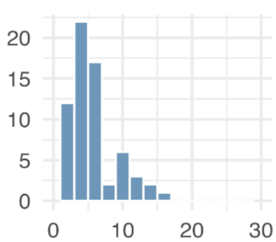
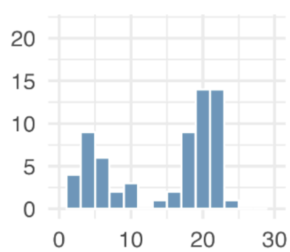
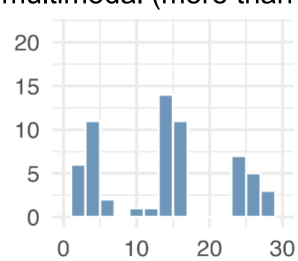
Explanatory variable – suspected cause (independent variable)
Response variable – suspected effect (dependent variable)

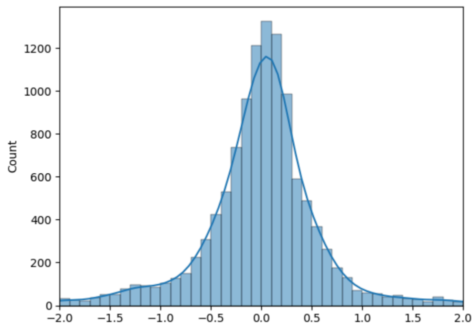
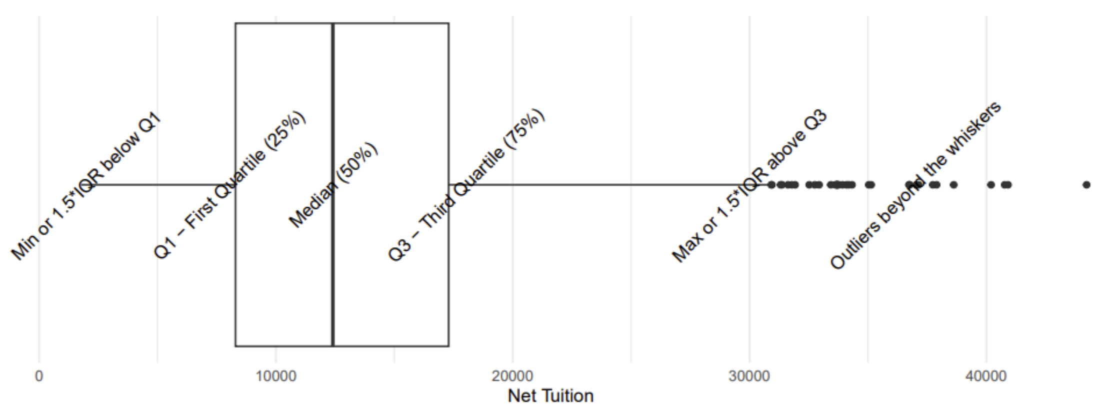
Univariate graph – show the distribution of a single variable.
Bivariate graph – show the relationship between two variables.

One Categorical

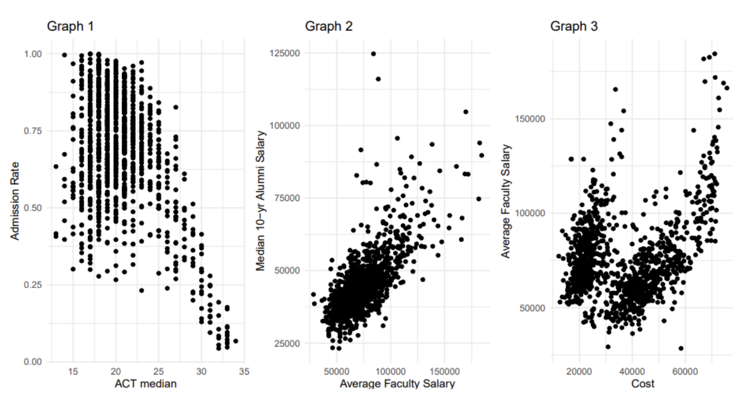
<table>
<tr>
<td>Bar chart</td>
<td>Vertical Bars

</td>
<td>Horizontal Bars

</td>
</tr>
</table>

One Quantitative

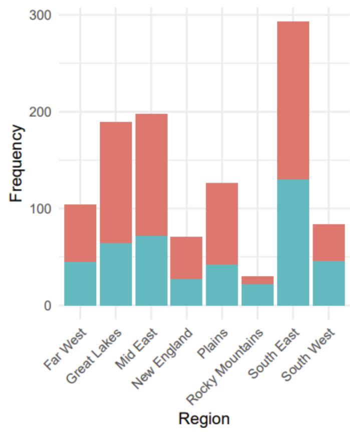
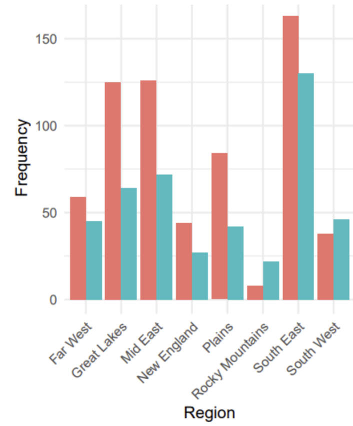
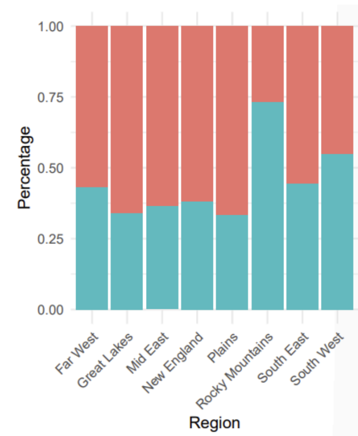
<table>
<tr>
<td rowspan="3">Histogram
Groups numeric values into equally spaced intervals known as bins.</td>
<td>Using 5 Bins

</td>
<td>Using 20 Bins

</td>
<td>Skewed Right

</td>
</tr>
<tr>
<td colspan="2">Skewed Left

</td>
<td>Symmetric

</td>
</tr>
<tr>
<td>Unimodal

</td>
<td>bimodal

</td>
<td>multimodal (more than two peaks)

</td>
</tr>
</table>

Density plot a smoothed out histogram	
Box Plot	**percentile** α – a number such that α% of our (quantitative) observations fall below this number. $Interquartile\ range = Q_3 - Q_1$ $Outlier = 1.5 \times IQR + Q_3$ $= Q_1 - 1.5 \times IQR$ 

Two Quantitative

Scatter Plot	

Two Categorical

Box Plot	Stacked	Clustered	Conditional
			

One quantitative variable:

1. Shape – is the distribution symmetric, skewed, bell-shaped, bimodal
2. Center – where are the data centered (mean and median)
3. Variability – How spread out are the data (range)
4. Unusual Points – outliers or excessive zeros

Two quantitative Variables:

1. Form – what type of trend or pattern exist(linear, non-linear, exponential, etc)
2. Strength – how closely do the data adhere to a trend or pattern (strong, moderate, weak)
3. Direction – how the values of one variable relate to the values of another variable(positive, negative)
4. Unusual Points – outliers or excessive zeros

Numerical Summaries

robust statistic – statistics that tends to not be influenced by outliers

– Measures of centrality

Mean ($\bar{X}$)	the arithmetic average of a variable (**not robust**)	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	the middle value of the data if arranged from smallest to largest. (**robust**)	

– Measures of Spread

Standard deviation(σ)	the average deviation(distance) of individual observations from the mean value. (**Not robust**)	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

		n = number of observation
Variance(σ^2)	measure that quantifies how much a set of numbers deviate from their mean.	$variance = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
Range	the difference between smallest and largest values	
Interquartile Range	the difference between the 75th quartile and the 25th quartile. (**Robust**)	$Interquartile\ range = Q_3 - Q_1$

Tables and Odds

– Tables

Contingency table – two way table in which both categorical variables have a binary response

	Event	Non-Event
Exposure	A	B
No Exposure	C	D

Exposure – presence of a factor that is being studied to determine its effect on a particular outcome
No Exposure – absence of a factor that is being studied to determine its effect on a particular outcome

Event – the occurrence of the outcome of interest
Non-Event – outcome did not occur

– Odds

$$Probability = \frac{number\ of\ success}{total\ number}$$

$$Odds = number\ of\ success : number\ of\ fail = \frac{number\ of\ success}{total\ number - number\ of\ success}$$

$$Odds\ Ratio = \frac{Odds\ of\ event\ in\ the\ exposed\ group}{Odds\ of\ event\ in\ the\ unexposed\ group}$$

OR = 1	No association between the exposure and the outcome.
OR > 1	Positive association – the exposure increases the likelihood of the event. Example: OR = 2, the event is twice as likely to occur in the exposed group as in the unexposed group.
OR < 1	Negative association –the exposure decreases the likelihood of the event.

	Example: OR = 0.5, the event is half as likely to occur in the exposed group as in the unexposed group.

Z-Score

z-scores describes a value's relationship to the mean of a group of values.

The Z score of an observation is defined as the number of standard deviations it falls above or below the mean.

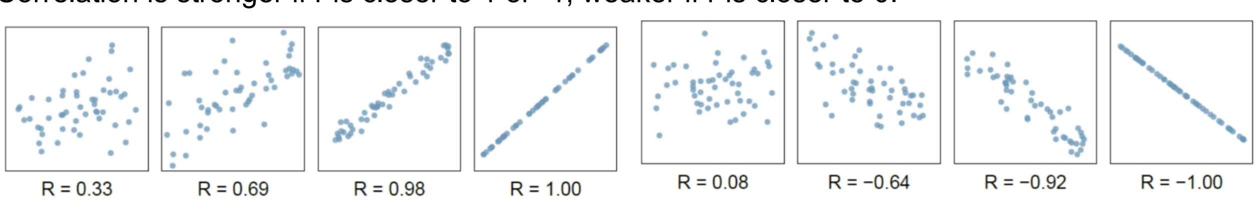
Z score of X_1 is Z_1 and Z score of X_2 is Z_2. If $|Z_1| > |Z_2|$, then X_1 is more unusual than X_2.

Examples	**Equation**
Observation is one standard deviation above the mean, Z score = 1. Observation is 1.5 standard deviations below the mean, Z score = -1.5.	$Z_i = \frac{x_i - \overline{x}}{S_x}$ x_i = one quantitative variable $\overline{x}$ = mean value of the variable S_x = standard deviations above/below the average

Regression

– Correlation

Correlation is stronger if r is closer to 1 or -1, weaker if r is closer to 0.



Pearson's correlation coefficient measures the strength of linear association between two quantitative variables	$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$ $= \frac{1}{n-1}\sum_{i=1}^{n}(z_{x_i})(z_{y_i})$
Spearman's rank correlation	

measures the strength of monotonic (non-linear) association between two quantitative variables	
Ecological correlations compare variables for data that have been aggregated at an ecological level. A correlation between two variables that are group means	**Ecological fallacy** – Ecological fallacies occur when we try to draw conclusions about individuals based on data collected at the group level.

– Regression line

Quantitative Regression	**Binary Categorical Regression**
$y = \beta_0 + x\beta_1$ β_0 = intercept β_1 = slope $\hat{y}$	$reference\ type = \beta_0 + 1_1\beta_1 + 0_2\beta_2$ β_1, β_2 = Coefficient $1_1, 0_2$ = Indicator Variables β_0 = Average of the reference type $\beta_1 + \beta_0$ = average for one indicator variable $\beta_2 + \beta_0$ = average for the other indicator variable

– Coefficient of determination R^2

Total sum of squares (SST)	distance between the mean of the observed and actual	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
Residual Sum of squares (SSR)	distance between the expected and actual	$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Coefficient of determination	proportion of the variance for a response variable that is explained by one or more explanatory variables	$R^2 = 1 - \frac{SSR}{SST} = r^2$

R^2 range from 0 to 1

- $R^2 = 1$ → regression model perfectly explains all the variability of the response variable
- $R^2 = 0$ → regression model explains none of the variability of the explanatory variables

Example:
$R^2 = 0.8$ → 80% of the variability in the response variable is explained by the explanatory variables, while the remaining 20% is unexplained and may be due to other factors or random noise.