

Today's Agenda

Exploring Dataset 1

1. Review the descriptive statistics
2. Build univariate visualizations

Justin Leinaweaver (Spring 2022)

Dataset 1: The Motivating Problem

What drives economic investment in US states?

Why do some states attract greater investment by companies and individuals than others?

Dataset 1: Let's Develop a Model

- 1 Literature Review
- 2 Exploratory Data Analysis
 - Descriptive Statistics
 - Univariate Visualizations

Dataset 1: Descriptive Statistics

Predictors to Analyze (5): Min wage, unemployment, population, homeowner rate and manufacturing

Mean	= AVERAGE
Median	= MEDIAN
Standard deviation	= STDEV.S
Minimum	= MIN
Maximum	= MAX
25th Percentile	= QUARTILE.EXC (<i>data</i> , 1)
75th Percentile	= QUARTILE.EXC (<i>data</i> , 3)

Dataset 1: Descriptive Statistics

Variable	mean	sd
homeowner_rate	69.1	5.1
manufacturing_thousands	241.9	243.2
min_wage	9.3	1.9
population_thousands	6575.4	7403.7
unemployment	7.4	1.8

Variable	Minimum	pct25	median	pct75	Maximum
homeowner_rate	53.6	67	69.7	72.4	78.2
manufacturing_thousands	9.5	59	167.7	328.5	1261.7
min_wage	7.2	7	9	11	13.5
population_thousands	582.3	1855	4561.3	7625.6	39368.1
unemployment	4.3	6	7.3	8.4	13

Variable	mean	sd	Minimum	pct25	median	pct75	Maximum
homeowner_rate	69.1	5.1	53.6	67	69.7	72.4	78.2
manufacturing_thousands	241.9	243.2	9.5	59	167.7	328.5	1261.7
min_wage	9.3	1.9	7.2	7	9	11	13.5
population_thousands	6575.4	7403.7	582.3	1855	4561.3	7625.6	39368.1
unemployment	7.4	1.8	4.3	6	7.3	8.4	13

- 1 Identify the **THREE** predictors with the **LEAST** variation across the states.
- 2 Identify the **THREE** predictors with the **MOST** variation across the states.


Resources on Moodle

▼ Class Videos and Resources




 Gharani, L. (2016, Feb 9). 3 Tips for Impressive Excel Charts.



 How to make a bar plot in Excel 365



 How to make box plots in Excel 365




 How to make scatter plots in Excel 365



 Creating Simple OLS Regressions (Excel 365)



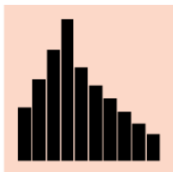
 Formatting an OLS Regression Table



Univariate Visualizations

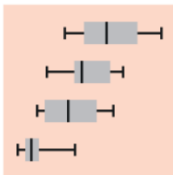
Numerical / Continuous Data

Histogram



The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Boxplot



Summarise multiple distributions by showing the median (centre) and range of the data

Categorical / Discrete Data

Column



The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

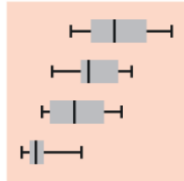


See above. Good when the data are not time series and labels have long category names.

Univariate Visualizations of GDP

Numerical / Continuous Data

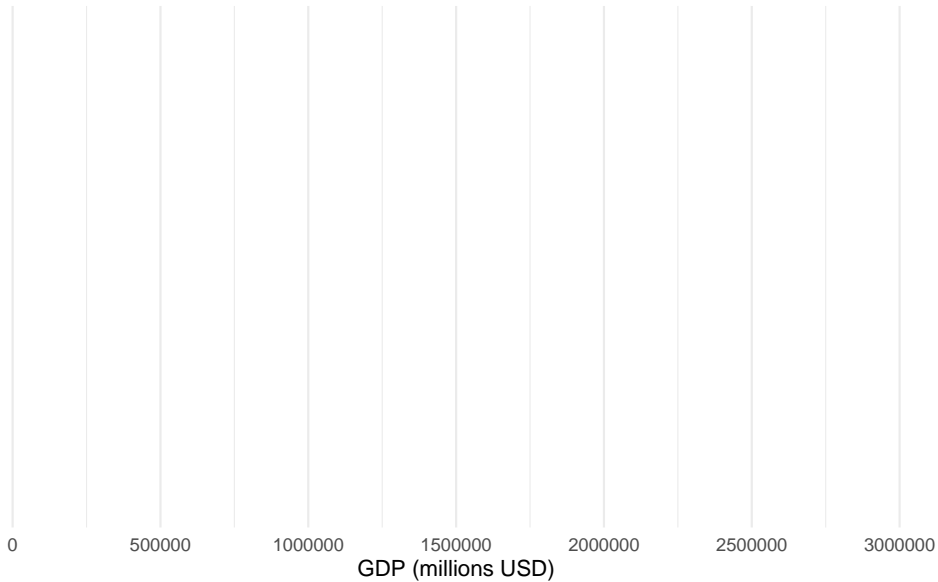
Boxplot



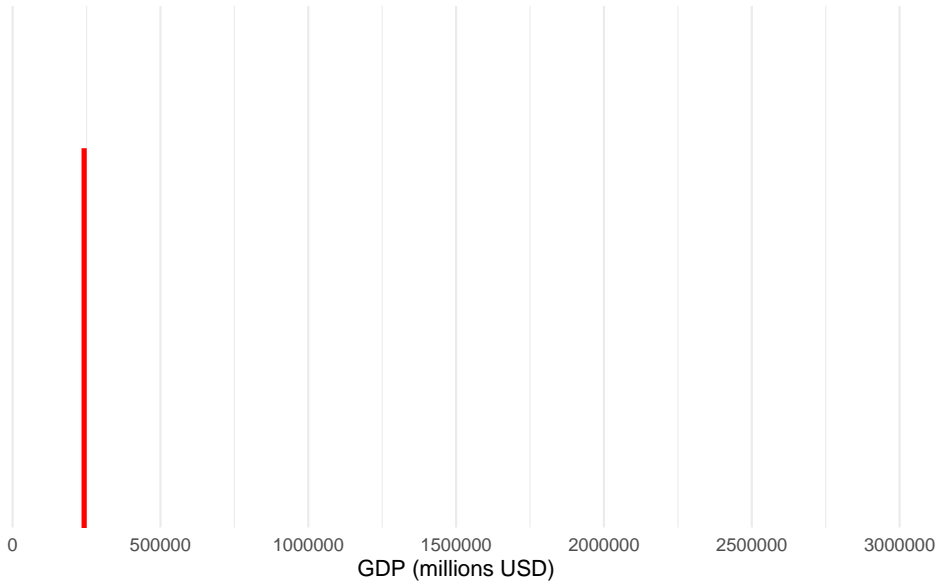
Summarise multiple distributions by showing the median (centre) and range of the data

Use **ONLY** the descriptive statistics for **GDP** to draw a boxplot **by hand**.

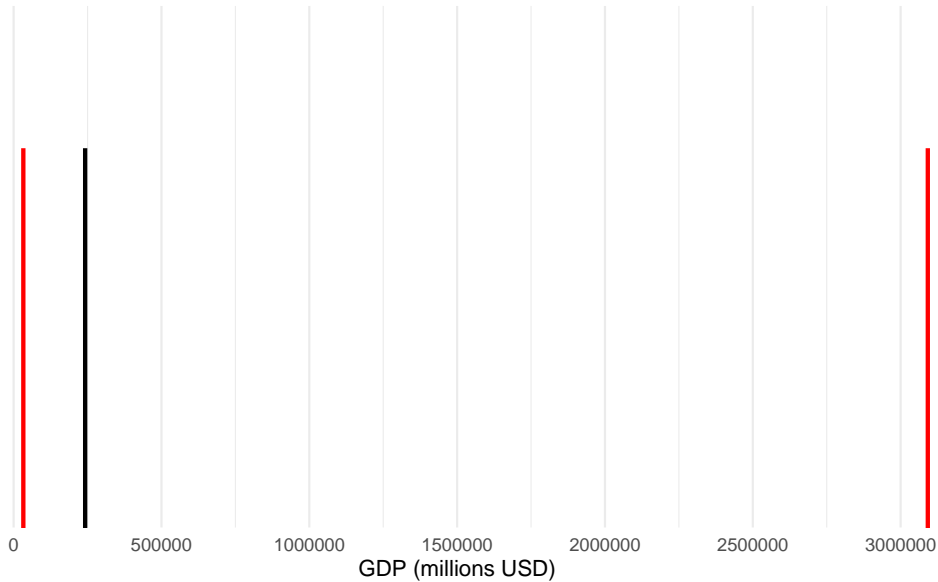
Set the X-Axis Range



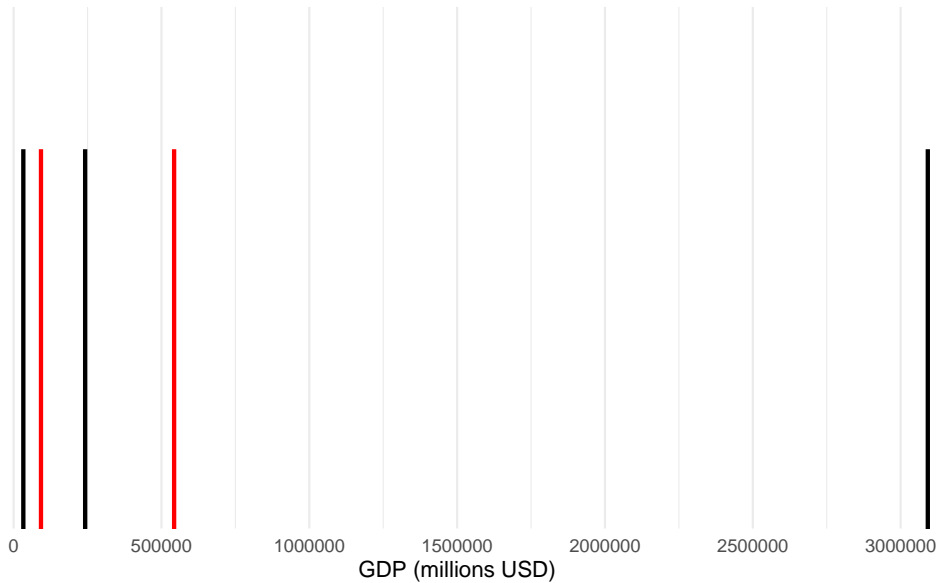
Mark the median (50th percentile)



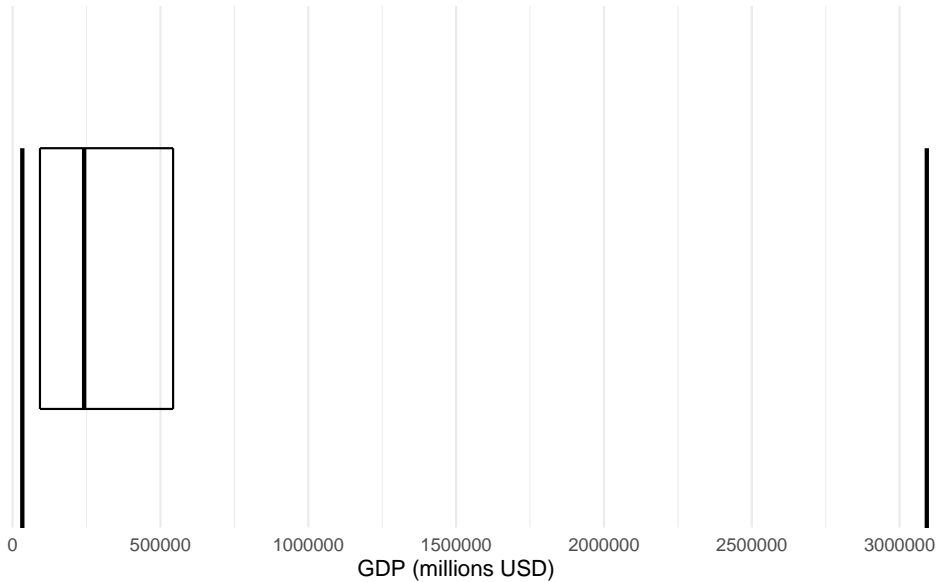
Add the minimum and maximum values



Add the 25th and 75th percentiles

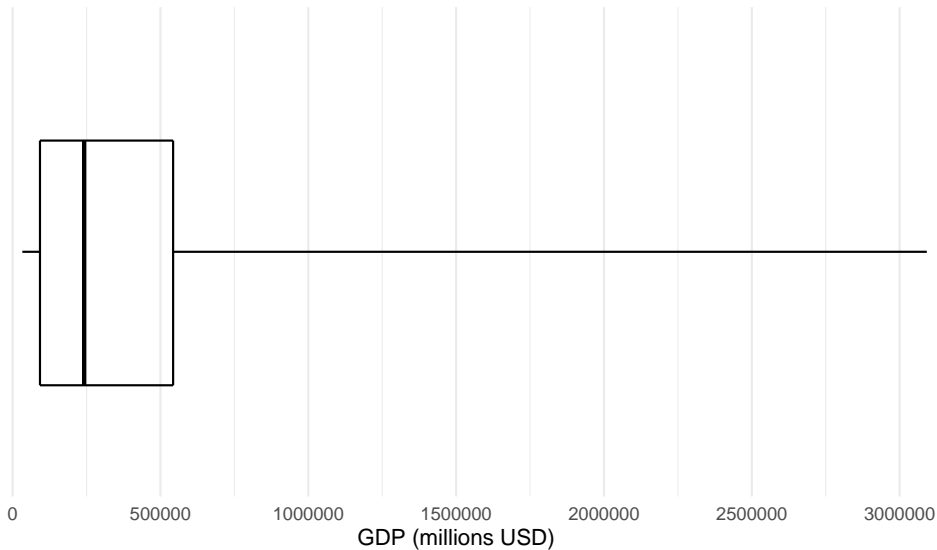


Replace IQR with a box



Replace min and max with whiskers

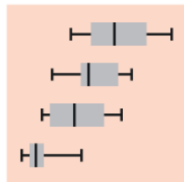
*Technically the whiskers should only extend 1.5x the IQR



Univariate Visualizations of GDP

Numerical / Continuous Data

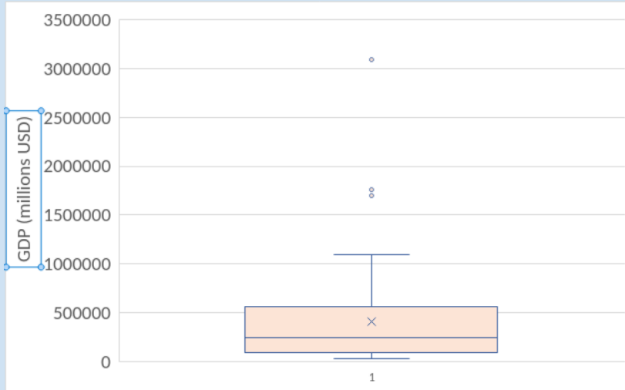
Boxplot



Summarise multiple distributions by showing the median (centre) and range of the data

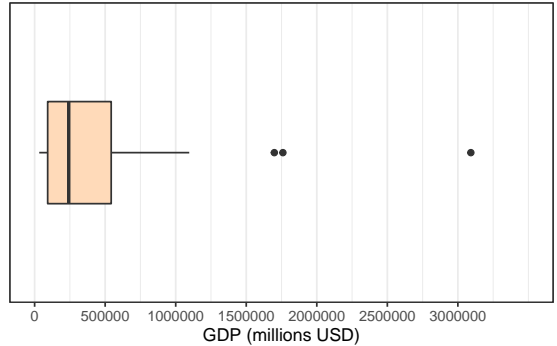
Use **Excel** to draw a boxplot of **GDP**.

Univariate Visualizations of GDP



Notes: Changed fill color, added an axis label and increased font size.

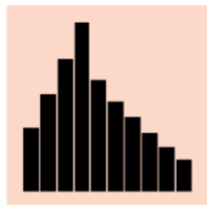
Univariate Visualizations of GDP



Univariate Visualizations of GDP

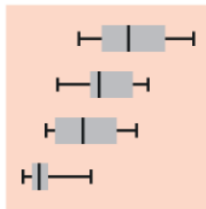
Numerical / Continuous Data

Histogram



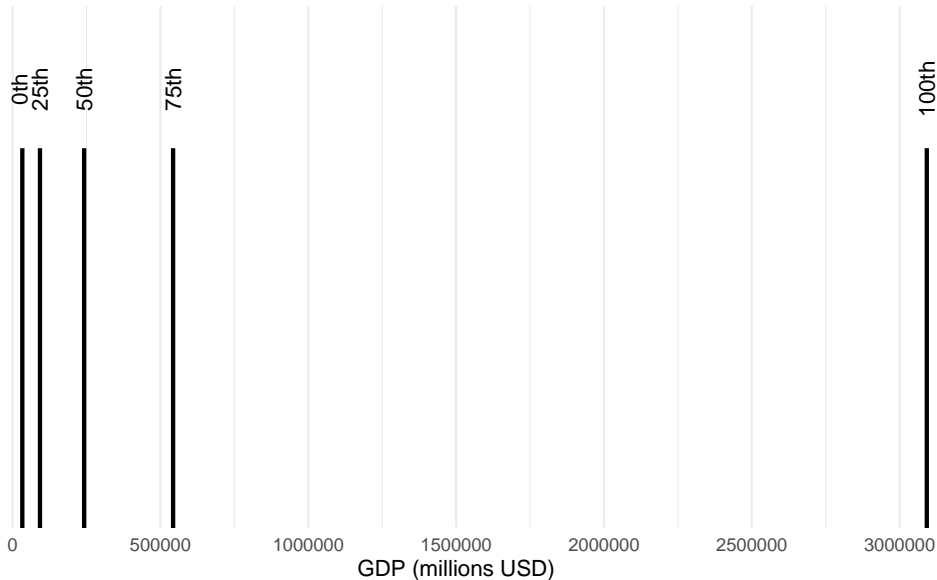
The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Boxplot



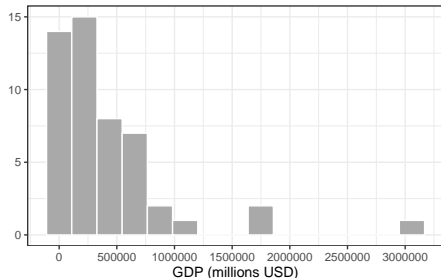
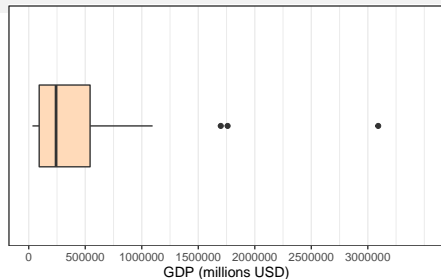
Summarise multiple distributions by showing the median (centre) and range of the data

Quartiles: One quarter of the data lies between each line (percentiles)



Univariate Analyses

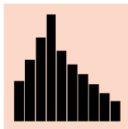
Statistic	GDP_millions
mean	413481.7
sd	537275.4
Minimum	32796.7
pct25	92470
median	241839.8
pct75	542802.3
Maximum	3091871.5



Univariate Visualizations of GDP Rate

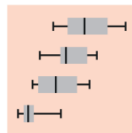
Numerical / Continuous Data

Histogram



The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Boxplot



Summarise multiple distributions by showing the median (centre) and range of the data

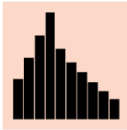
By hand, draw a histogram of **GDP rate**

Remember: Use the box plot to build your histogram

Univariate Visualizations of GDP Rate

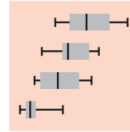
Numerical / Continuous Data

Histogram



The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Boxplot

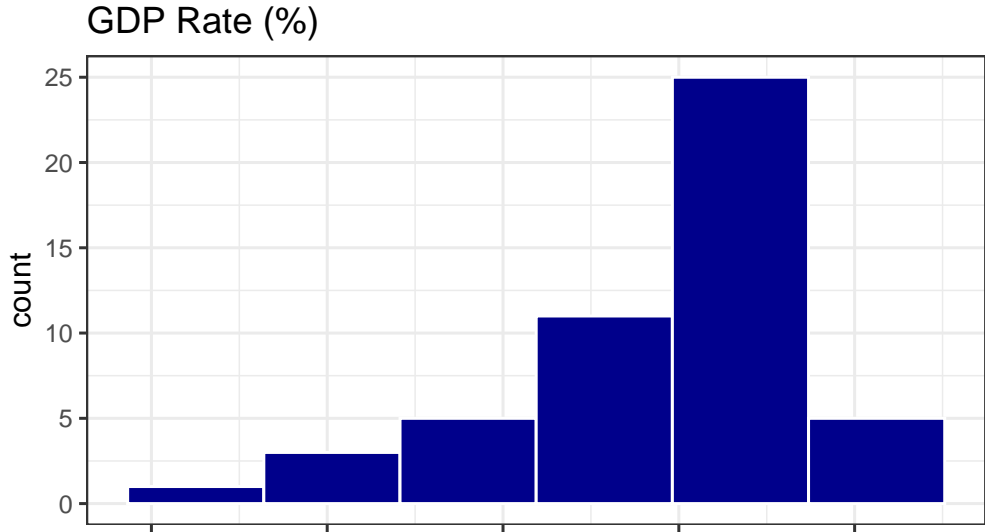


Summarise multiple distributions by showing the median (centre) and range of the data

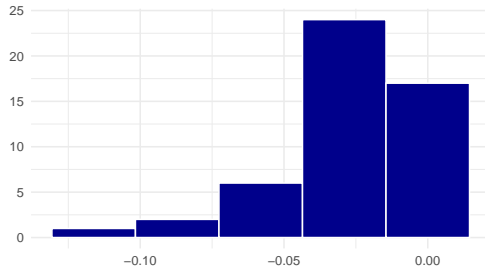
Use **Excel** to check your histogram of **GDP rate**

'Format Axis' and set the number of bins to 5

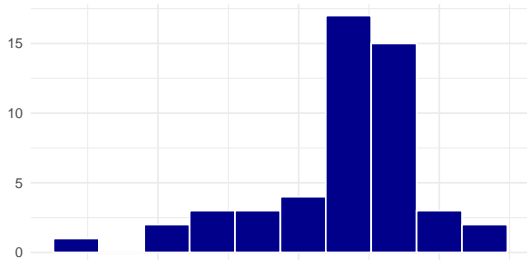
Univariate Visualizations of GDP Rate



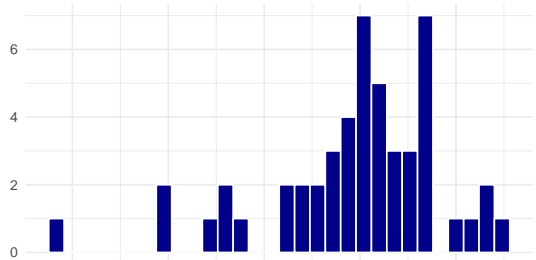
5 Bins



10 Bins



30 Bins



Univariate Viz of GDP Categories

Column



The standard way to compare the size of things. Must always start at 0 on the axis.

Bar



See above. Good when the data are not time series and labels have long category names.

Categorical / Discrete Data

By hand, make a bar plot of **gdp category**

Univariate Viz of GDP Categories

Column



The standard way to compare the size of things. Must always start at 0 on the axis.

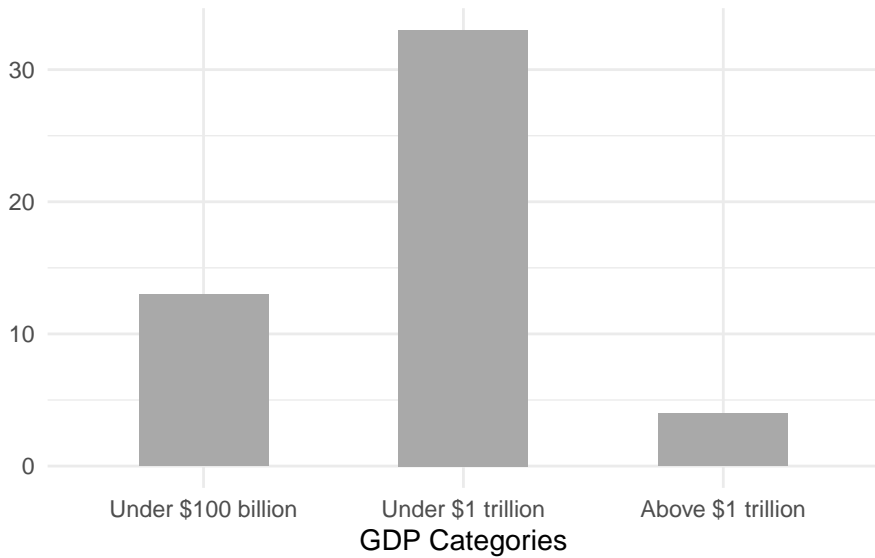
Bar



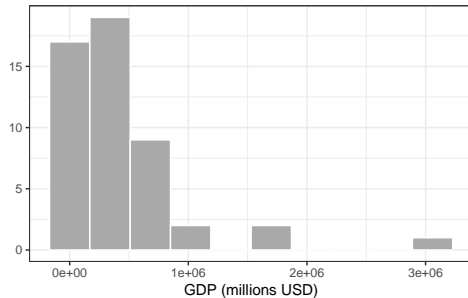
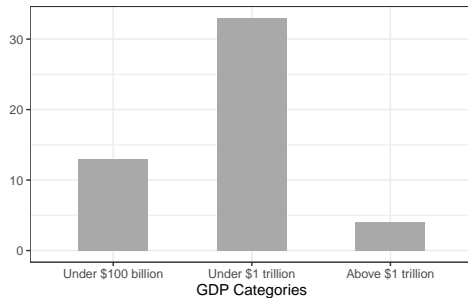
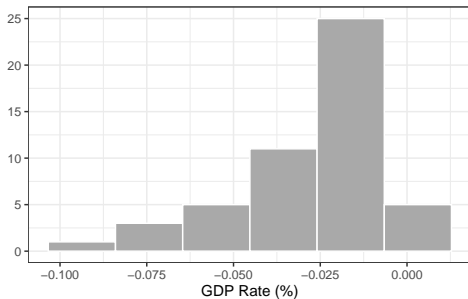
See above. Good when the data are not time series and labels have long category names.

Categorical / Discrete Data

Use **Excel** to check your bar plot of **gdp category**



Three ways to visualize GDP



Use Excel to make an appropriate univariate visualization for each of the remaining variables in the dataset.

Developing a Model: Univariate Analyses

- ① Which is more useful the descriptive statistics or the visualization?

Developing a Model: Univariate Analyses

- ① Which is more useful the descriptive statistics or the visualization?
- ② In what specific ways can we use univariate analyses to answer our motivating question?