Project Overview

In this project, you will use R and apply exploratory data analysis techniques to explore relationships in one variable to multiple variables and to explore a selected data set for distributions, outliers, and anomalies.

Prepare for this project with: **Data Analysis with R**.

Note

If you have successfully completed the project for the Data Analysis with R course in the past (which entails having graduated from the course and having access to your course certificate), simply email us at dataanalyst-project@udacity.com with your passing evaluation and we'll give you credit for this project.

What do I need to install?

In order to complete the project, you will need to install R. You can download and **install R from the Comprehensive R Archive Network (CRAN)**.

After installing R, you will need to download and install **R Studio**. Choose the appropriate installation for your operating system.

Finally, you will need to install a few packages. We recommend opening R Studio and installing the following packages using the command line.

```
install.packages("ggplot2", dependencies = T)
install.packages("knitr", dependencies = T)
install.packages("dplyr", dependencies = T)
```

For more information on installing R packages, please refer to **Installing R Packages** on R Bloggers.

Why this Project?

MENTORSHIP

EDA can lead to insights, which may uncover to other questions, and eventually predictive models. It also is an important “line of defense” against bad data and is an opportunity to notice that your assumptions or intuitions about a data set are violated.

What will I learn?

After completing the project, you will:

- Understand the distribution of a variable and to check for anomalies and outliers
- Learn how to quantify and visualize individual variables within a data set by using appropriate plots such as scatter plots, histograms, bar charts, and box plots
- Explore variables to identify the most important variables and relationships within a data set before building predictive models; calculate correlations, and investigate conditional means
- Learn powerful methods and visualizations for examining relationships among multiple variables, such as reshaping data frames and using aesthetics like color and shape to uncover more information

Why is this Important to my Career?

In this project, you learn skills to frame and present data. Data, by itself, is "ubiquitous and cheap," says Google's Chief Economist and UC Berkeley professor Hal Varian. What you do as a data analyst is take that data and turn it into insights.

When working on this project, think about the insights you're trying to provide. When speaking to an employer, this is what they want to first hear about.

NEXT