Data Science Orientation

Lab – Exploring Data

Overview

Rosie Reeves is an entrepreneurial middle-school student who sells homemade cold drinks from a stand during the summer months. This summer, Rosie sold two flavors of drink (lemon and orange) at two different locations (the beach and the park); and to promote her drinks stall, she distributed leaflets in the local area. Rosie recorded details of her sales and leaflet distribution in a text file, along with a note of the temperature each day.

In this lab, you will explore and visualize the data Rosie recorded.

Note: The figures used in the lab files for this lab are not the same as the figures used in the course demonstrations!

What You'll Need

To complete the labs, you will need the following:

- A Windows or Mac OS X computer running Excel 2016. If you do not have Excel 2016 installed, you can sign up for a free trial of Office 365 at http://aka.ms/edx-dat101x-o365, and install Office applications on your Windows PC or Mac.
- The lab files for this course. You can download and extract these from http://aka.ms/edx-dat101x-labs.

Exercise 1: Exploring Data in Excel

In this exercise, you will use Excel to explore and visualize the data that Rosie had collected.

Import the Data into Excel

1. Create a new blank Excel workbook and import the text data from the **Lemonade2016.csv** comma-delimited text file in the **DAT101x_Labfiles** folder where you extracted the lab files.
2. Format the imported data as a table.

Cleanse the Data

1. Find and remove any duplicate rows in the data.
2. Find and resolve any missing values in the data, either by interpolating missing values that are in an obvious sequence, or by entering the average value for the column (rounded to the nearest whole number) into any cells where a value is missing.

Add Derived Columns

1. Add a column named **Sales** to the table, in which the total sales of lemon and orange is calculated.
2. Add a column named **Revenue** to the table, in which the sales revenue is calculated by multiplying the total sales and price.
3. In an empty cell below the table, insert a formula to calculate the total revenue for the summer, and make a note of this value.

Find the Highest and Lowest Temperatures

1. Use conditional formatting to identify the top 10% and bottom 10% temperatures.
2. Note the highest and lowest temperatures.

Create Charts

1. Create a line chart that shows **Date** and **Revenue**. Then add a linear trendline to determine whether the revenue trend is rising, falling, or staying level.
2. Create a scatter-plot chart that shows **Leaflets** on the X axis and **Sales** on the Y axis, and determine whether any signs of linear correlation can be detected.
3. Create a histogram that shows **Revenue** distributed into 10 bins, and note whether the distribution of this data is normal, left skewed, or right skewed.

Exercise 2: Using Statistics in Excel

In this exercise, you will use the Data Analysis Pack in Excel to apply some statistical functions to Rosie's sales data.

Calculate Descriptive Statistics

1. Calculate descriptive statistics for the **Temperature**, **Leaflets**, **Price**, and **Sales** columns.
2. Note the *mean*, *median*, and *mode* for each column.
3. Note the *range*, *sample variance*, and *standard deviation* for each column.

Calculate Correlation

1. Determine the strength and direction of any correlations between the **Temperature**, **Leaflets**, and **Price** columns and the **Sales** column.

Compare Sales of Lemon and Orange Flavors

1. Perform an appropriate test to determine whether there is a statistically significant difference between sales of **Lemon** and **Orange** drinks.

Perform Regression

1. Perform regression using **Sales** as the Y input and **Temperature**, **Leaflets**, and **Price** as the X input.
2. Review the results of the regression, noting the relative significance of the input variables and the intercept and weights applied to them to calculate the predicted value for Y.
3. Use the intercept and weights to calculate a predicted Y value for the following features:
    - **Temperature**: 80
    - **Leaflets**: 110
    - **Price**: 0.35

4. Calculate predicted sales for the above scenario if the number of leaflets distributed is increased to 120.