Lab 8. Forecasting with Tableau

In this lab, we will cover the following recipes:

- Basic forecasting and statistical inference
- Forecasting on a dataset with outliers

Technical requirements

To follow the recipes from this lab, you will need to have Tableau 2020.1 installed.

In the following recipes, we will be using the `hormonal_response_to_excercise.csv` and `stock_prices.csv` datasets, which you can download from the following URLs:

- https://github.com/SlavenRB/Forecasting-with-Tableau/blob/master/hormonal_response_to_excercise.csv
- https://github.com/SlavenRB/Forecasting-with-Tableau/blob/master/stock-prices.txt Please make sure you have a local copy of the dataset saved to your device before we begin.

Basic forecasting and statistical inference

The aim of this recipe is to introduce a basic forecasting method that relies on linear regression. We are going to use a built-in Tableau facility for linear regression. Simply put, regression analysis helps us discover predictors of a variable that we are interested in. We model the relationship between potential predictors and our variable of interest. Once we establish the model of the relationship between predictors and our variable, we can use it for further predictions.

To perform for casting, we will use the `hormonal_response_to_excercise.csv` dataset. This dataset comes from a health behavior study that aimed to explore the factors influencing cortisol response while exerting the maximal, peak effort during physical exercise (the `Cortmax` variable in our dataset).

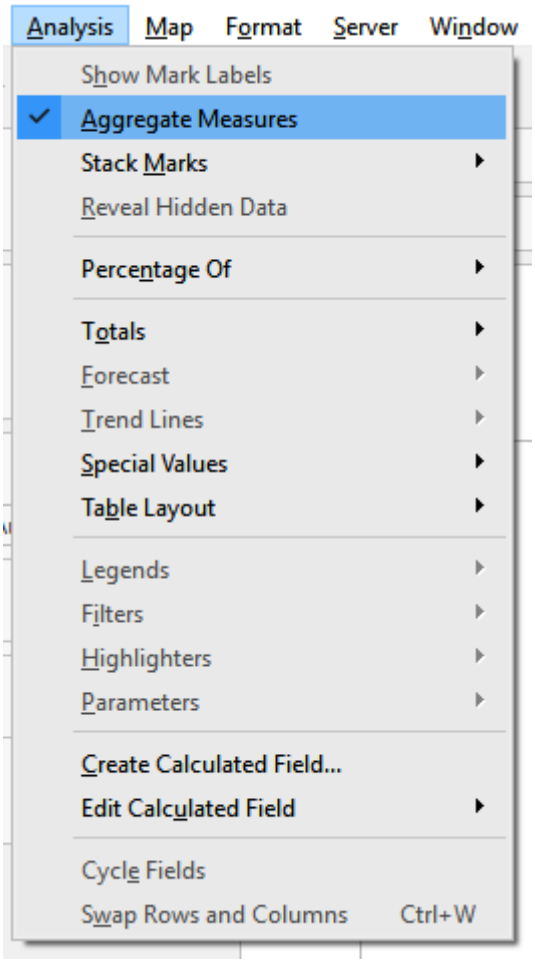
Our first task is to explore how effectively we can predict the level of cortisol response in the point of maximal effort during physical exercise, based on the cortisol level at rest (the `Cortrest` variable). So, in this example, our variable of interest (variable we are trying to predict) is `Cortmax` (cortisol level during maximal effort), while our predictor variable (the one that we are using to make a prediction) is `Cortrest` (cortisol level during rest). We will try to model the relationship between these two variables.

Getting ready

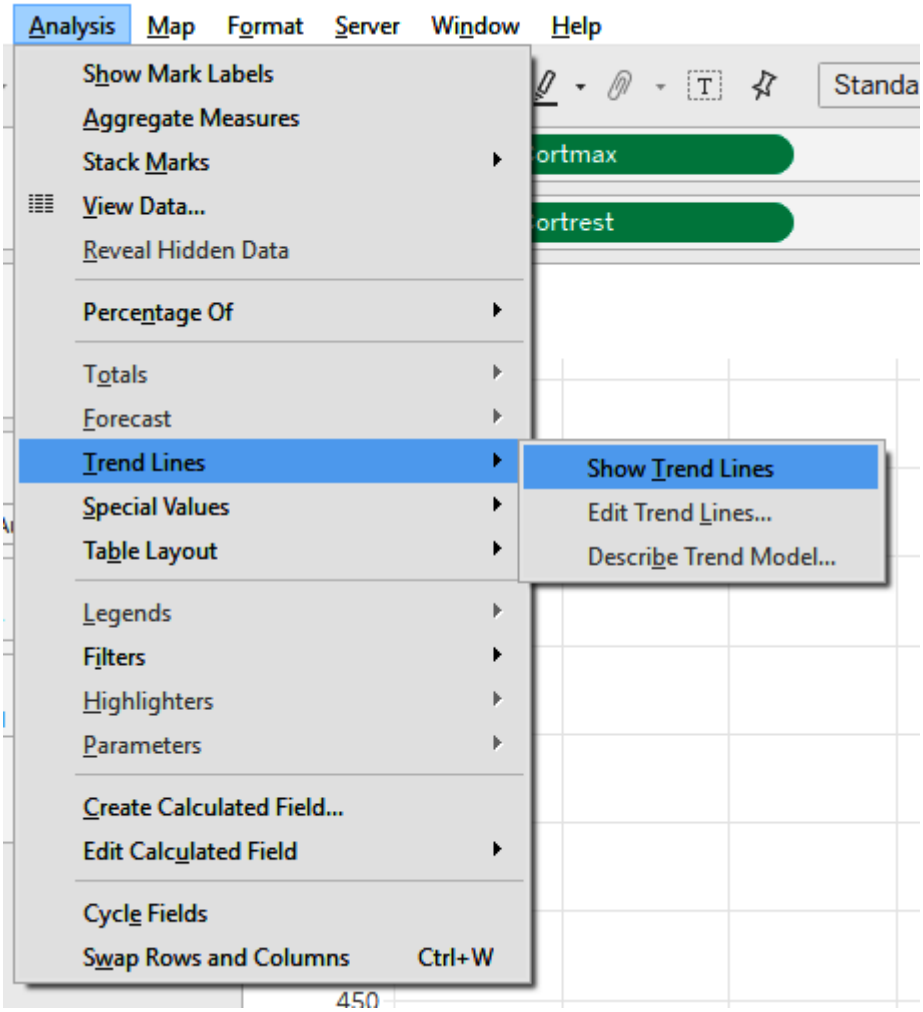
To perform the steps outlined In this lab, you will need to connect to the `hormonal_response_to_excercise.csv` dataset.

How to do it...

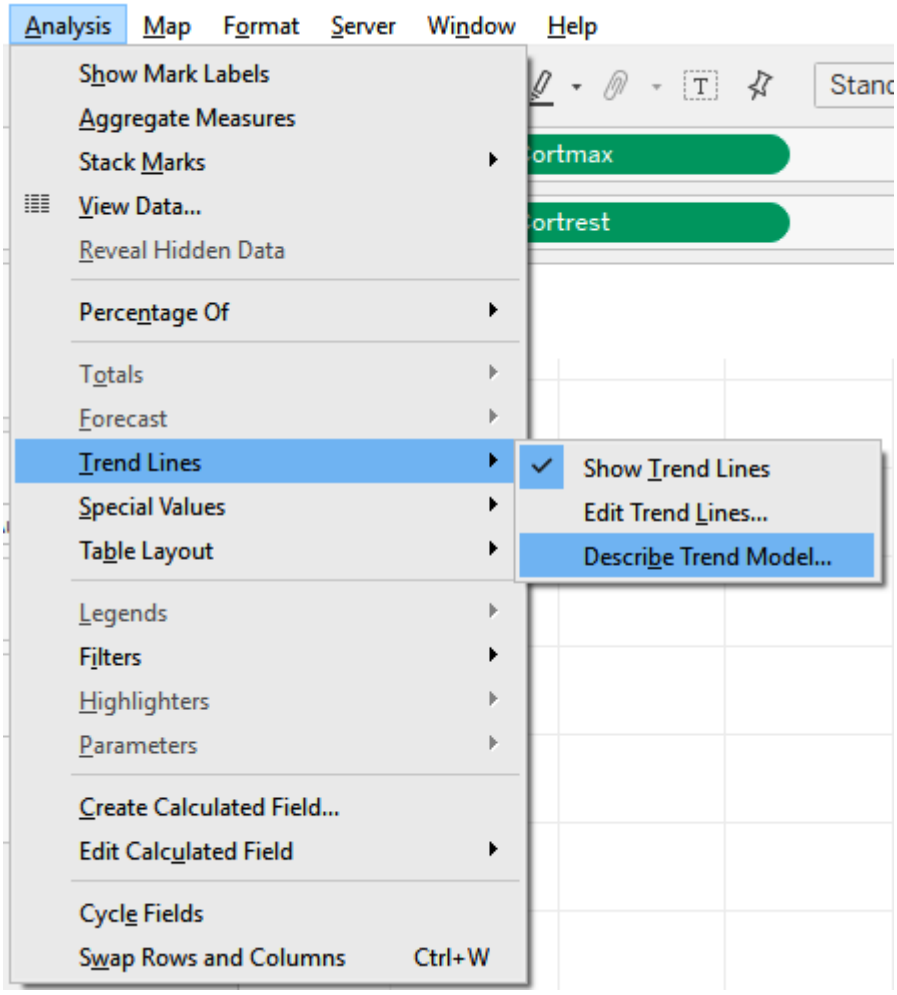
1. Open a blank worksheet and drag and drop **`Cortmax`** from **`Measures`** into the **`Columns`** shelf.
2. Drag and drop **`Cortrest`** from **`Measures`** to the **`Rows`** shelf.
3. In the main menu toolbar, navigate to ** `Analysis` **and in the drop-down menu deselect **`Aggregate Measures`**:



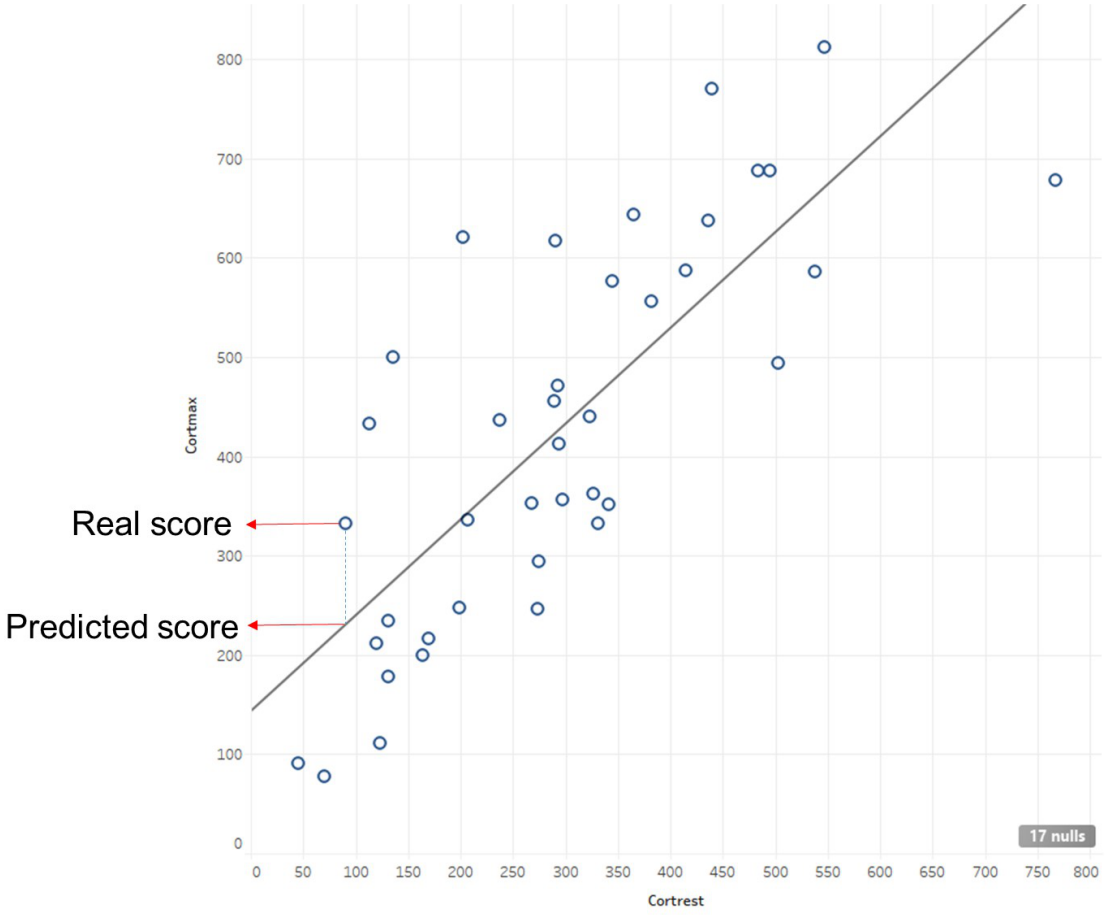
4. In the main menu toolbar, navigate to ** `Analysis` **and in the drop-down menu, under **`Trend Lines`**, select **`Show Trend Lines`**:



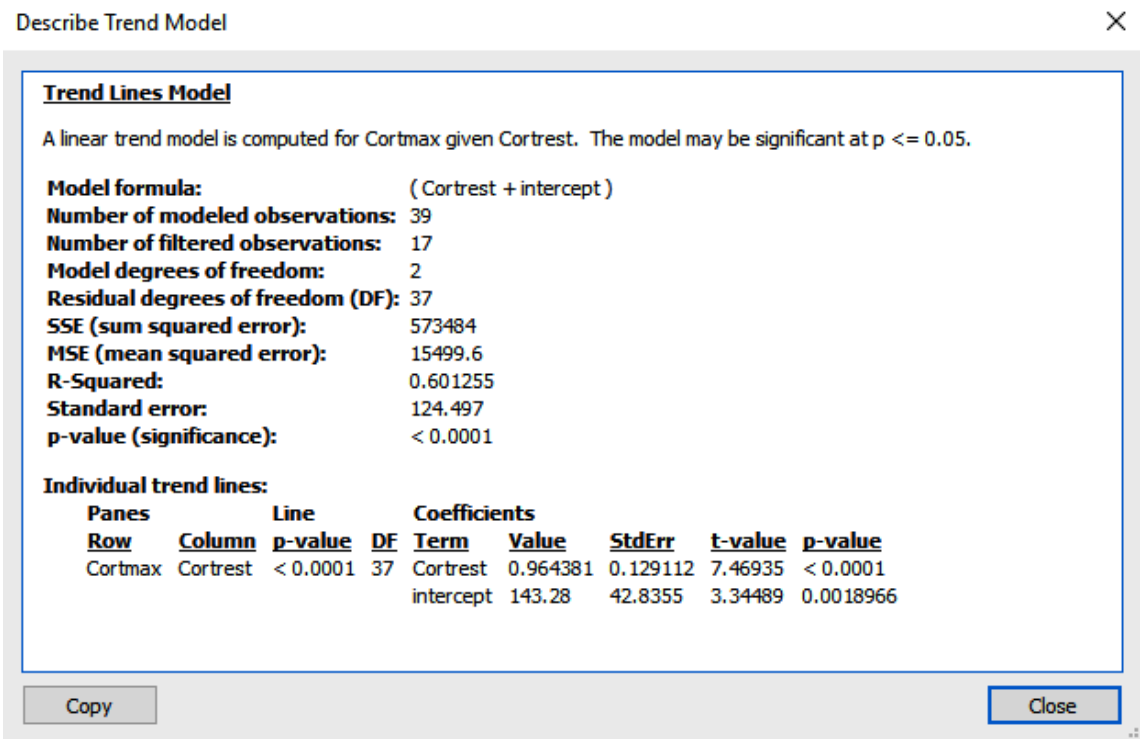
5. Once again, navigate to `Trend Lines` **under** `Analysis` in the main menu toolbar, and select `Describe Trend Model...`:



The following is the output:



In the following screenshot, we can see that we have successfully modeled the relationship between these two variables:



How it works...

We have created a regression model. Now, we are going to interpret the results, which tell us how successful our model is. But first, we need to get familiar with some basic statistics. In the broadest terms, the aim of each model is to represent real-life phenomena. All models differ in accuracy, or how well they depict reality. In statistics, we call the accuracy of the model its fit. The fit of a model is better if the difference between real data (that we measured) and predicted data (based on our model) is smaller.

We tried to predict cortisol level during maximum effort based on cortisol level during rest. Actual data points are represented with the circles, and the predictions that we made based on our model are vertically projected on the line (shown in the previous screenshot). As you can see, some circles lie almost on the line, some are above, and some are below the line. But the line is positioned so that these differences are minimized. If we want to estimate how good our model is, we should estimate the size of these differences. But, since the differences have both positive and negative values (some points are above and some are below the line), we can not just simply sum it up because they would cancel out. That's why we first need to square each difference and then sum it up. The result is [**sum squared error**] ([**SSE**])---that is a measure of our model's error, or how much it deviates from actual data. In order to estimate the goodness of our model's fit, we need to compare the size of that deviation with a benchmark. The most commonly used benchmark is the simple average of [*y*] value or the flat, horizontal line. Comparison of our model and baseline gives us R-squared. The bigger the R-squared is, the smaller the probability that we obtained it by chance. The conventionally accepted threshold of the probability is 0.05 and is denoted by the p-value (significance) in our output. If our p-value is smaller than the threshold, we can conclude that we have enough evidence to believe that our model is good enough. In our case, the p-value is much smaller than the mentioned threshold, so we can assume that the cortisol level during the maximum effort can be reasonably well predicted based on the cortisol level at rest. We can conclude that we have created a successful model of the relationship between these two variables.

There's more...

The regression that we presented in this example is linear because we assumed that the relationship between our variables was linear -- an assumption that turned out to be correct. However, other types of models are also available in Tableau. They can be accessed by navigating to **`Analysis`** | **`Trend Lines`** | **`Edit Trend Lines...`**. In the **`Trend Lines Options`** window, we can choose other models such as **`Logarithmic`**, **`Exponential`**, **`Power`**, or **`Polynomial`**. A good way to choose the model is to first plot the data and visually inspect it.

Forecasting on a dataset with outliers

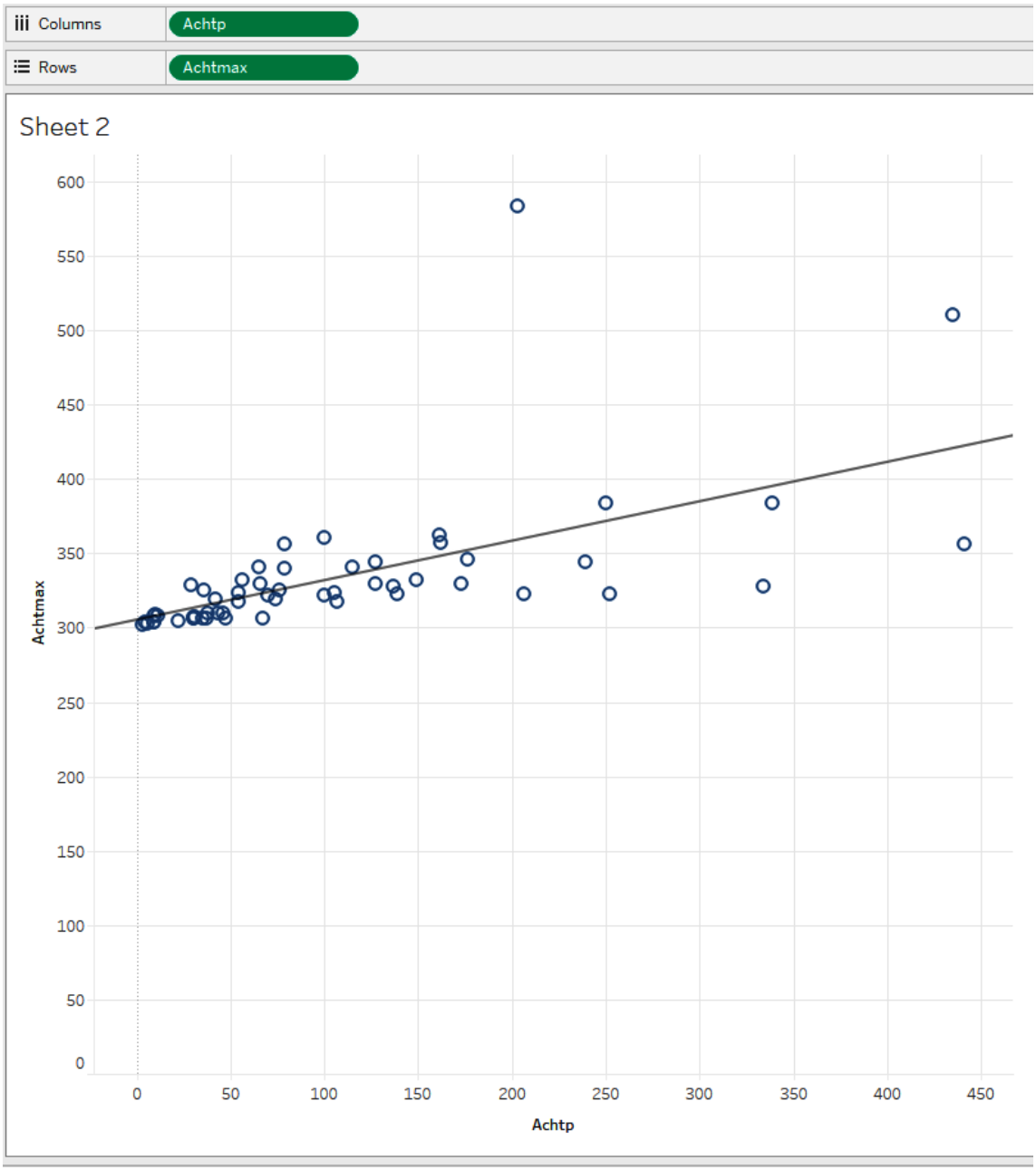
In this lab, we are going to learn how to deal with outliers. Outliers are data points that are very unusual, atypical, and deviate from the trend present in the majority of the dataset. Outliers can be dangerous if not dealt with appropriately because they can significantly skew the results of an analysis. In this recipe, we will explore ways of detecting outliers in Tableau. We are going to perform a regression analysis and see how the regression line is affected by these cases.

Getting ready

For this recipe, we need the `hormonal_response_to_excercise.csv` dataset. We are going to use the `Achtp` and `Achtmax` variables. The `Achtp` variable is the level of adrenocorticotropic hormone at the beginning of the test, while `Achtmax` is the level of adrenocorticotropic hormone at the maximum effort during physical exertion.

How to do it...

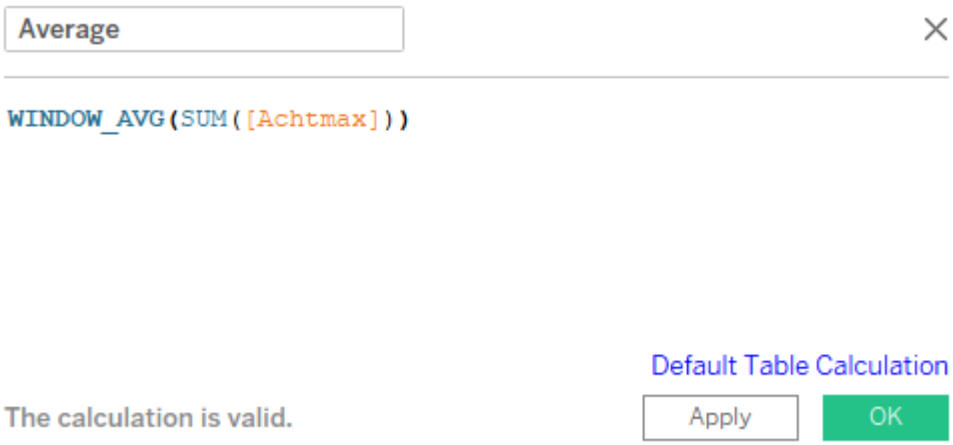
1. Drag and drop **`Achtp`** from **`Measures`** into the **`Columns`** shelf.
2. Drag and drop **`Achtmax`** from **`Measures`** into the **`Rows`** shelf.
3. In the main menu toolbar, in the **`Analysis`** drop-down menu, deselect **`Aggregate Measures`**.
4. In the main menu toolbar, in the** `Analysis` ** drop-down menu, navigate to **`Trend Lines`** | **`Show Trend Lines`**:



5. Rename the sheet to `Outliers included`.
6. In the main menu navigate to **`Analysis`** | **`Create Calculated Field...`**.
7. Rename the calculated field from** `Calculation 1` ** to `Average`, and in the formula space, type the following expression:

```
WINDOW_AVG(SUM([Achtmax]))
```

The calculation field shows the preceding expression in the following screenshot:



8. Click on **`OK`** to save and exit the calculated field editor window.
9. Repeat [*step 7*] to create another calculated field. Name the field `Lower` and in the formula space type the following expression:

```
[Average] - 2.5*WINDOW_STDEV(SUM([Achtmax]))
```

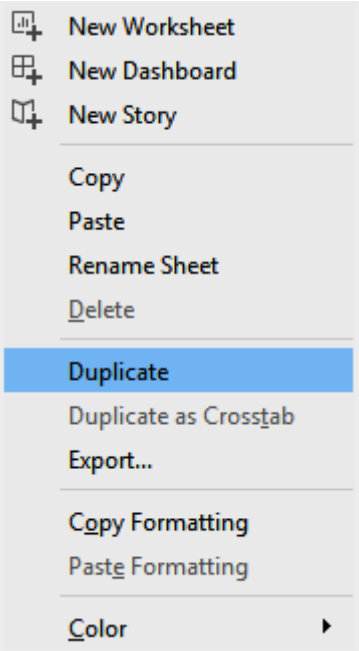
10. Save and exit by clicking **`OK`**.
11. Repeat [*step 7*] to create yet another calculated field. Name the field `Upper` and in the formula space, type the following expression:

```
[Average] + 2.5*WINDOW_STDEV(SUM([Achtmax]))
```

12. Save and exit by clicking on **`OK`**.
13. Repeat [*step 7*] one last time to create our final calculated field. Name this field `Outliers` and in the formula space, type the following expression:

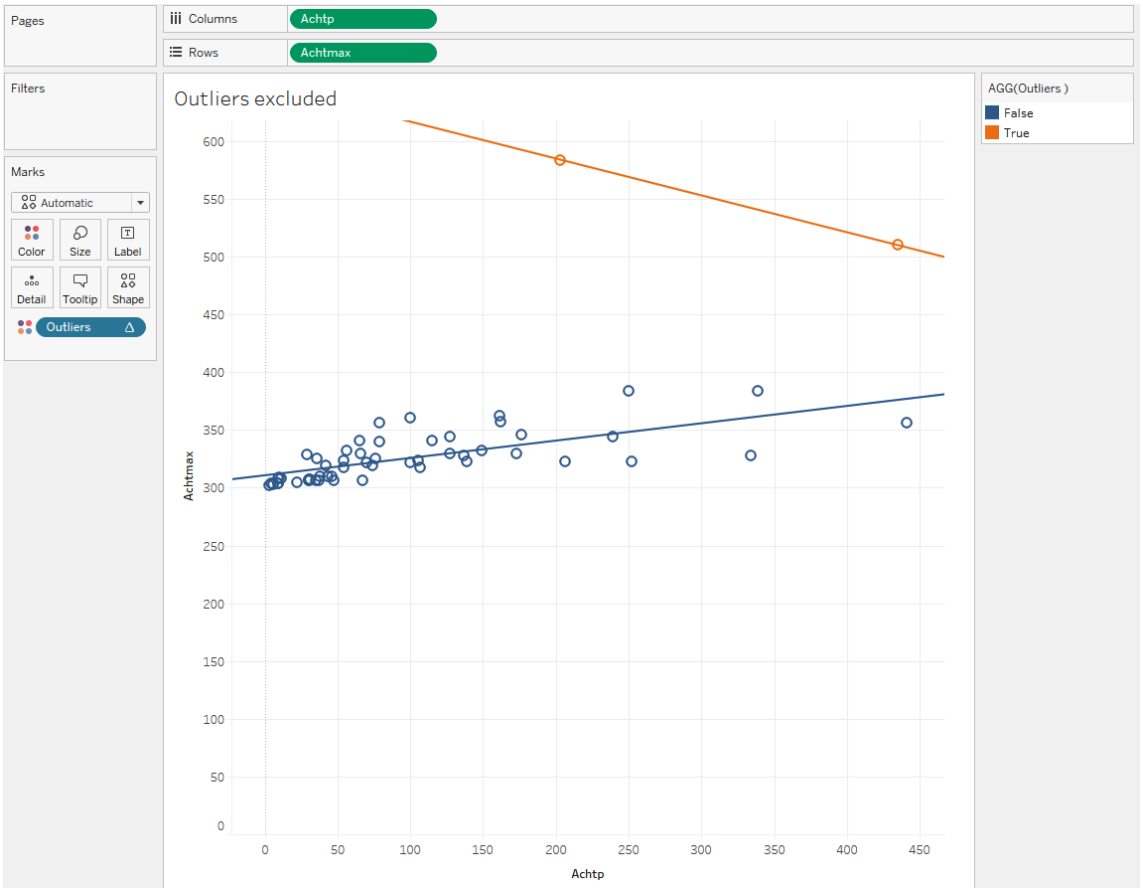
```
SUM([Achtmax])> [Upper] or SUM([Achtmax]) < [Lower]
```

14. Save and exit by clicking on **`OK`**.
15. Right-click on the ** `Outliers included` **sheet tab at the bottom of the workspace and select **`Duplicate`** as seen in the following screenshot:

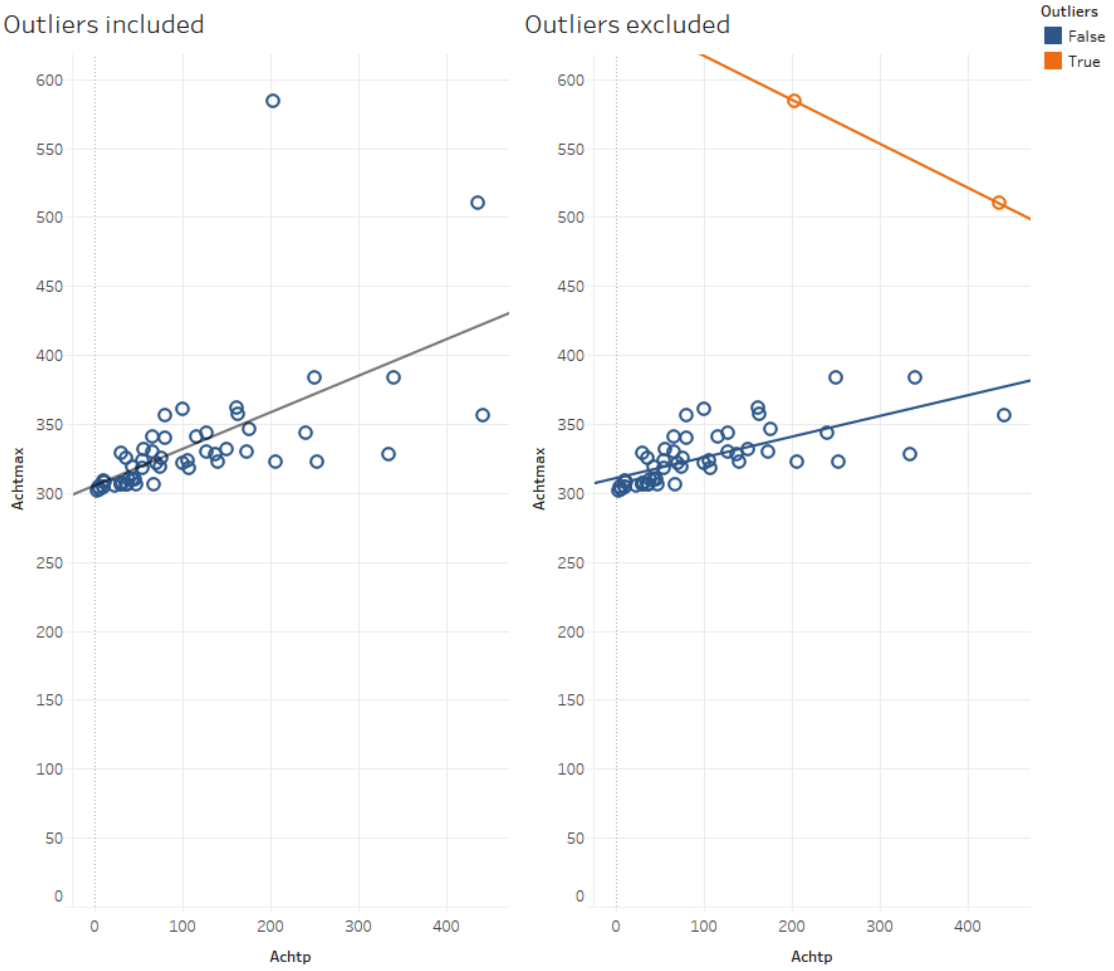


16. This will create an identical sheet named **`Outliers included (2)`**. Rename this sheet `Outliers excluded`.

17. Drag and drop **`Outliers`** from **`Measures`** to **`Color`** in the **`Marks`** card:



18. In the main menu toolbar, navigate to **`Dashboard`** | **`New Dashboard`** .
19. Drag and drop ** `Outliers included` **sheet from the **`Sheets`** section of the **`Dashboard`** pane into the canvas.
20. Drag and drop the **`Outliers excluded`** sheet from the **`Sheets`** section of the **`Dashboard`** pane into the canvas, to the right of the ** `Outliers included` **sheet in the chart:



How it works...

In this lab, we learned how to detect outliers. Outliers are extreme values that stand out from the other values in the sample. In order to detect outliers, we relied on a commonly used conventional rule---outliers are all values that deviate from the mean more than +/-2.5 standard deviations, which excludes around 1% of our sample.

In this example, we are able to see how outliers can influence statistical models. We can see that the model that includes outliers has a much steeper slope than the model that excludes outliers, meaning that they have pulled our linear model away from the majority of data, giving us skewed results. When interpreting results, we have to pay special attention to this (so-called leverage) effect. Otherwise, we risk declaring a statistical effect significant even when it does not actually exist.