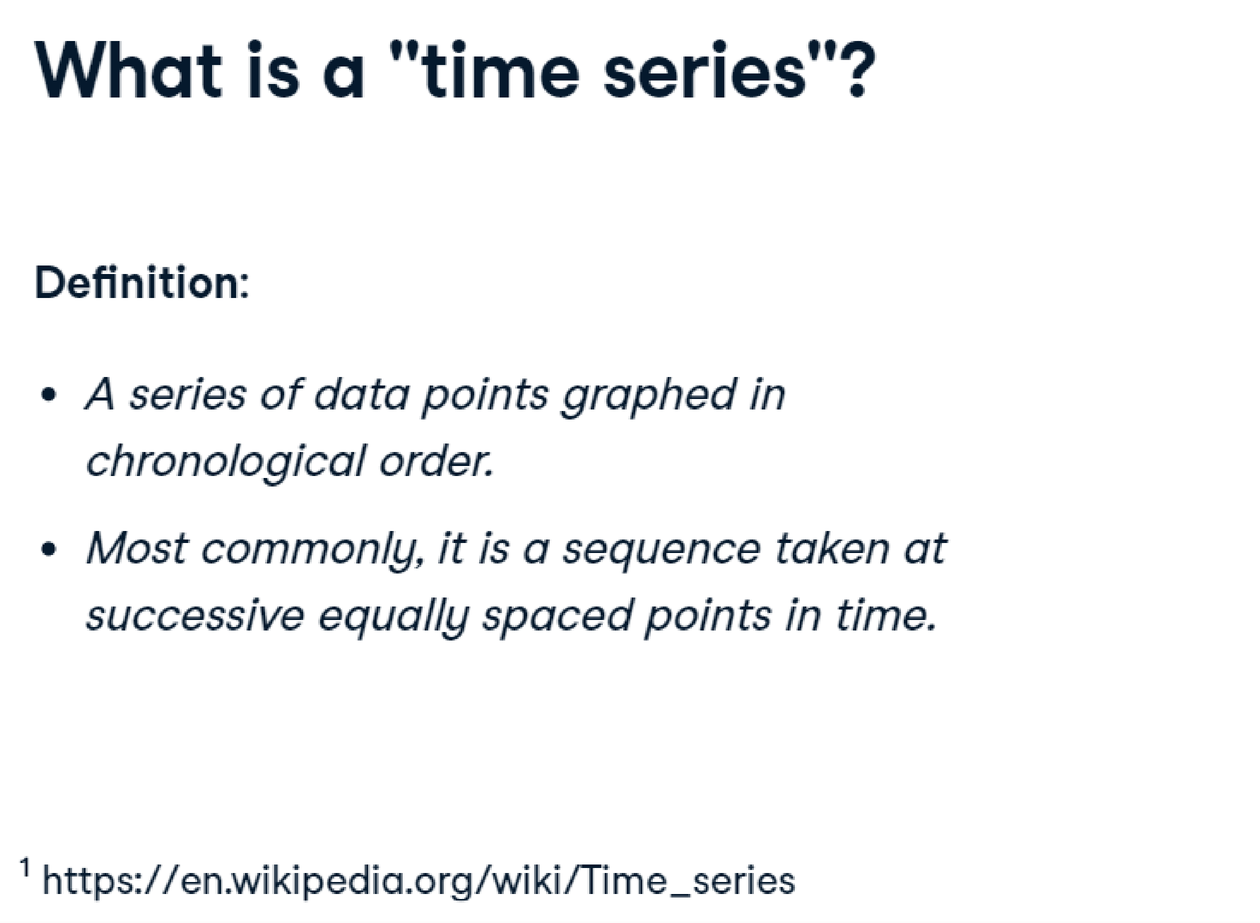
Trend Analysis in Power BI

LEARNING VOYAGE

Table of Contents

1: Exploring Time Series Data



What is a "time series"?

Definition:

- *A series of data points graphed in chronological order.*
- *Most commonly, it is a sequence taken at successive equally spaced points in time.*

[1] https://en.wikipedia.org/wiki/Time_series

Use cases for time series analysis

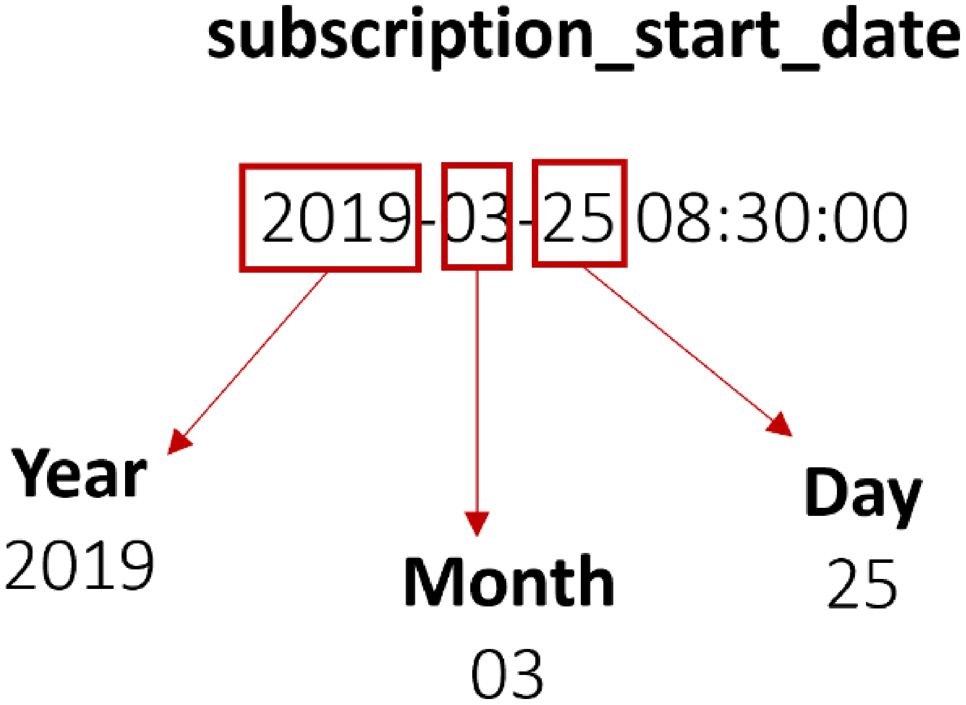
Use Cases:

- Patterns (e.g. cyclical) in a variable
- Season-specific trends
- Systemic challenges
- Relationships with a target outcome
- Informing a forecasting model

Examples:

- Weather
- Heart rate monitoring
- Quarterly sales
- Interest rates

Analyzing at different date grains



subscription_start_date

2019-03-25 08:30:00

Year

2019

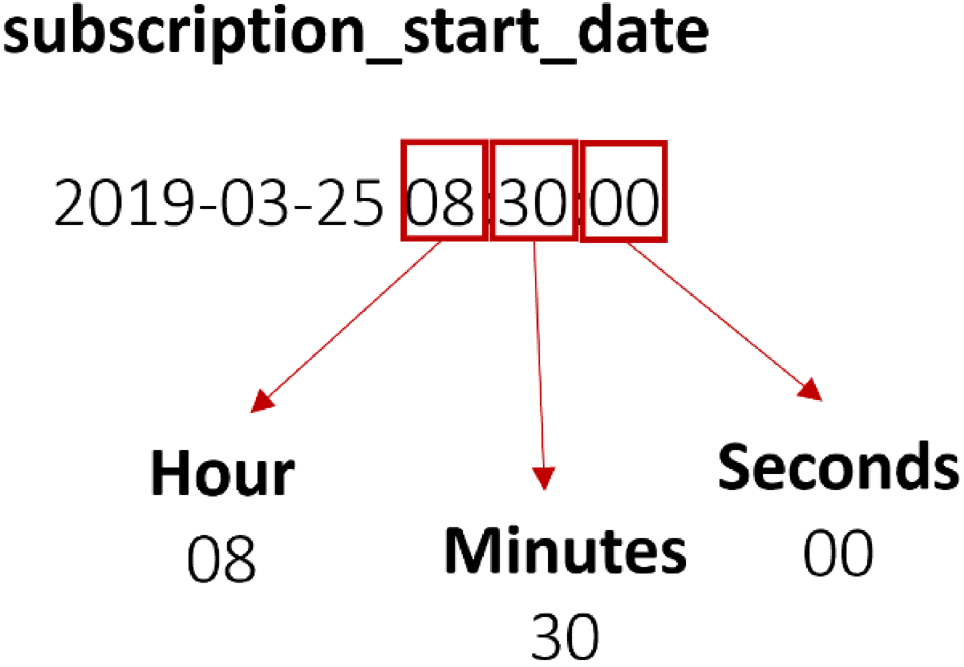
Month

03

Day

25

Analyzing at different date grains



subscription_start_date

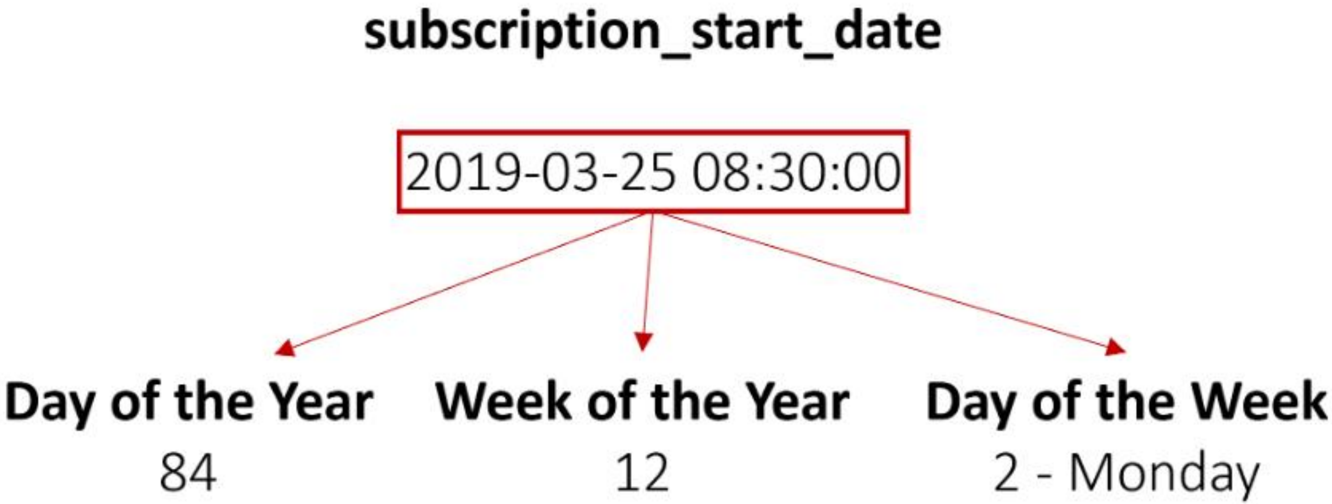
2019-03-25 08:30:00

Hour
08

Minutes
30

Seconds
00

Mutating into different time variables



subscription_start_date

2019-03-25 08:30:00

Day of the Year
84

Week of the Year
12

Day of the Week
2 - Monday

Mutating into different time variables

subscription_start_date

2019-03-25 08:30:00

age: 1,003 days

age_group:
old

current_date

2021-12-22 12:00:00

Measuring the change over a period of time

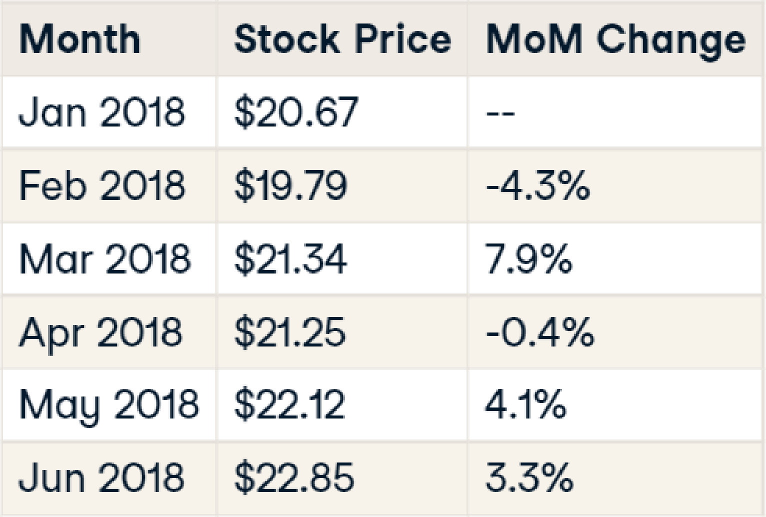
Period-over-period change

$$\frac{current\_period - previous\_period}{previous\_period}$$

Measuring the change over a period of time

Period-over-period change

$$\frac{\mathit{current\_period} - \mathit{previous\_period}}{\mathit{previous\_period}}$$



Month	Stock Price	MoM Change
Jan 2018	$20.67	--
Feb 2018	$19.79	-4.3%
Mar 2018	$21.34	7.9%
Apr 2018	$21.25	-0.4%
May 2018	$22.12	4.1%
Jun 2018	$22.85	3.3%

DAX functions for dates

`DATE()` : constructs a date value from parts (e.g. year, month, and day)

`LEFT()` : extracts a given number of characters from a set of characters, starting from the left side.

`RIGHT()` : extracts a given number of characters from a set of characters, starting from the right side.

`MID()` : extracts a given number of characters from a set of characters, starting from a defined place in the set.

`WEEKDAY()` : returns the day of the week as a number; default is to use the number 1 for Sunday.

Let's practice!

Statements about time series analysis
Which statement about time series analysis is false?

Possible Answers

○ It involves a sequence of data points in chronological order.

○ It is the only mechanism to forecast future trends or events.

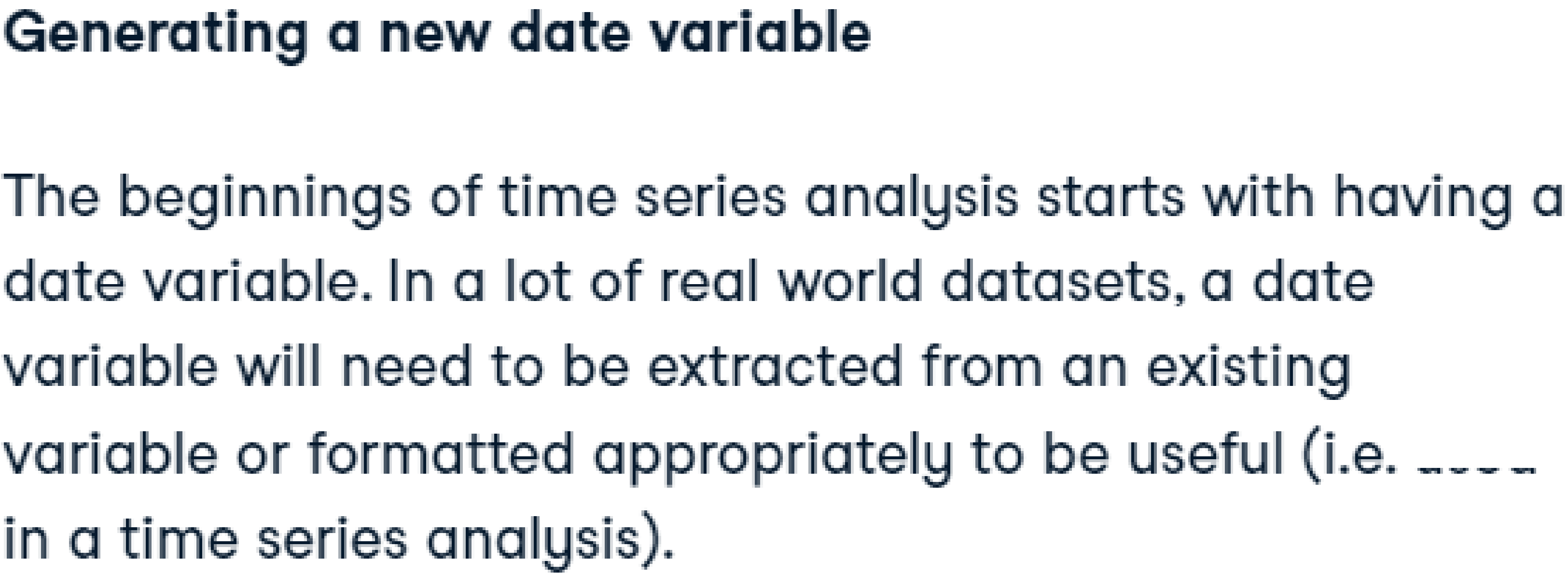
○ It is used to understand patterns within a target variable.

○ It can be used to determine seasonal trends of book sales.

DEMO

Exploring AirBnB time series

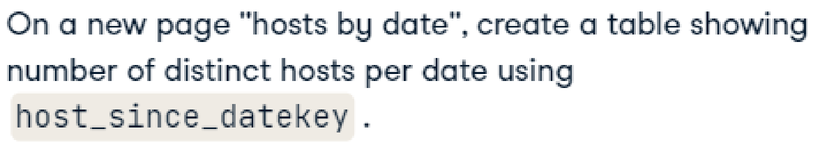
"Complete Lab"



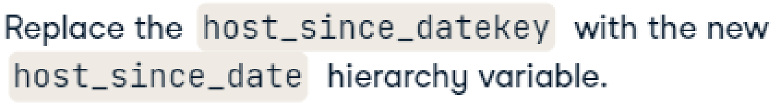
Generating a new date variable

The beginnings of time series analysis starts with having a date variable. In a lot of real world datasets, a date variable will need to be extracted from an existing variable or formatted appropriately to be useful (i.e. [illegible] in a time series analysis).

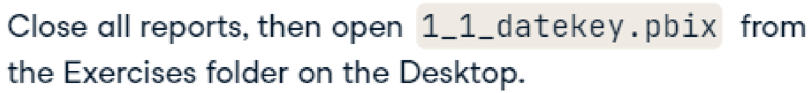
“Instructions"



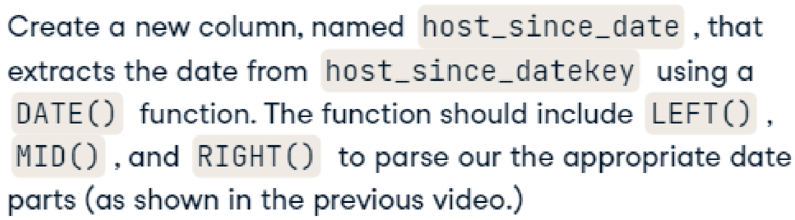
On a new page "hosts by date", create a table showing number of distinct hosts per date using `host_since_datekey` .



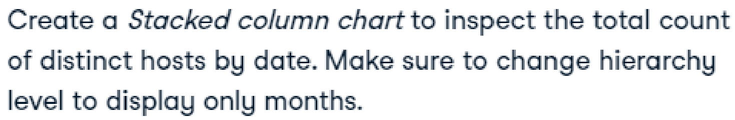
Replace the `host_since_datekey` with the new `host_since_date` hierarchy variable.



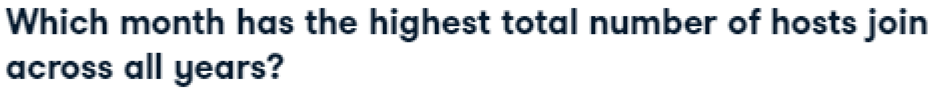
Close all reports, then open `1_1_datekey.pbix` from the Exercises folder on the Desktop.



Create a new column, named `host_since_date` , that extracts the date from `host_since_datekey` using a `DATE()` function. The function should include `LEFT()` , `MID()` , and `RIGHT()` to parse our the appropriate date parts (as shown in the previous video.)

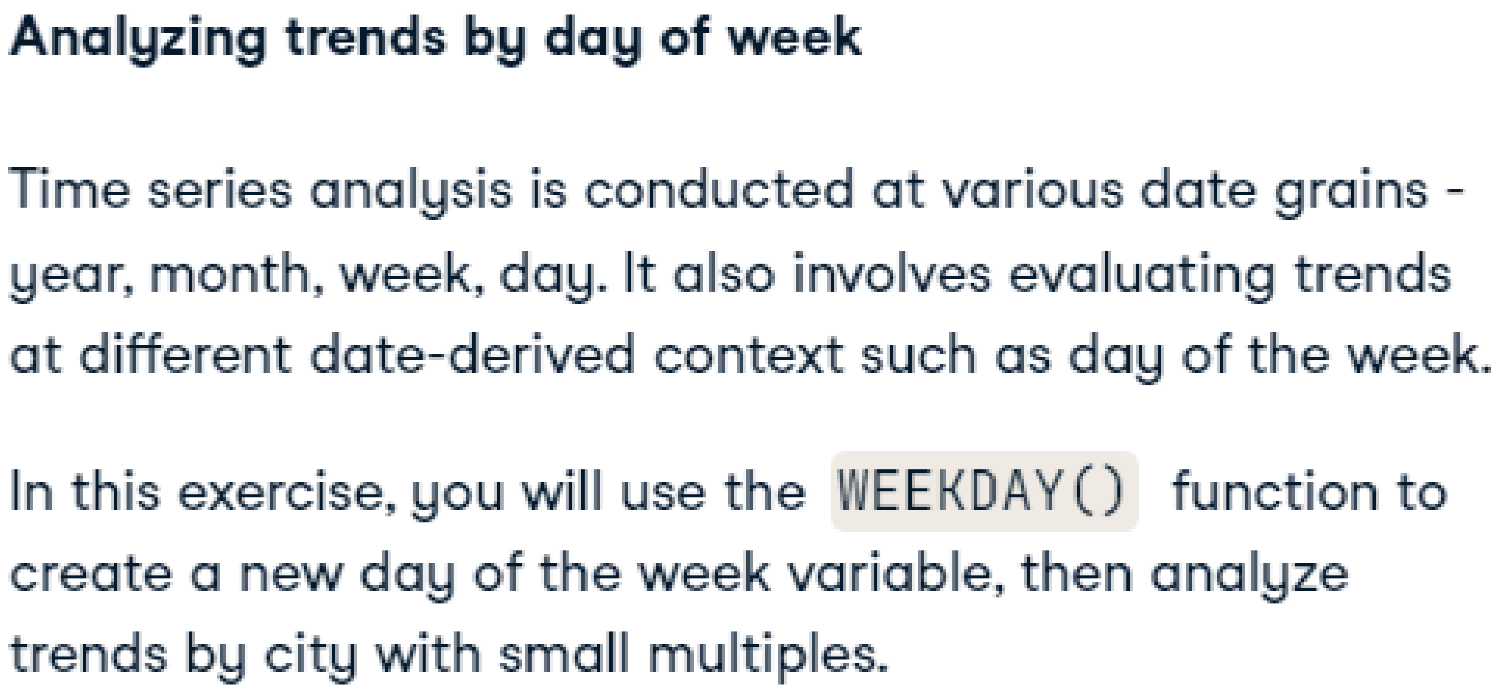


Create a *Stacked column chart* to inspect the total count of distinct hosts by date. Make sure to change hierarchy level to display only months.



Which month has the highest total number of hosts join across all years?

"Complete Exercise"

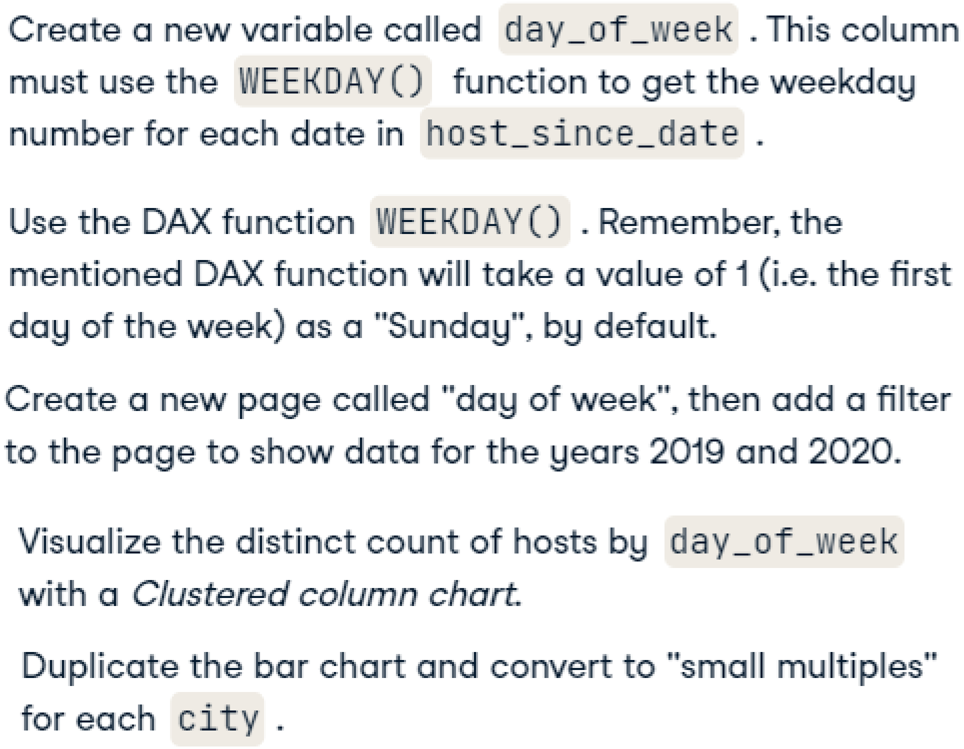


Analyzing trends by day of week

Time series analysis is conducted at various date grains - year, month, week, day. It also involves evaluating trends at different date-derived context such as day of the week.

In this exercise, you will use the `WEEKDAY()` function to create a new day of the week variable, then analyze trends by city with small multiples.

“Instructions"



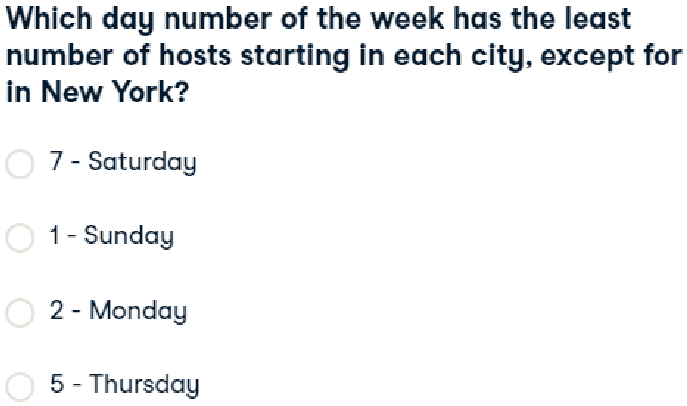
Create a new variable called `day_of_week` . This column must use the `WEEKDAY()` function to get the weekday number for each date in `host_since_date` .

Use the DAX function `WEEKDAY()` . Remember, the mentioned DAX function will take a value of 1 (i.e. the first day of the week) as a "Sunday", by default.

Create a new page called "day of week", then add a filter to the page to show data for the years 2019 and 2020.

Visualize the distinct count of hosts by `day_of_week` with a *Clustered column chart.*

Duplicate the bar chart and convert to "small multiples" for each `city` .



Which day number of the week has the least number of hosts starting in each city, except for in New York?

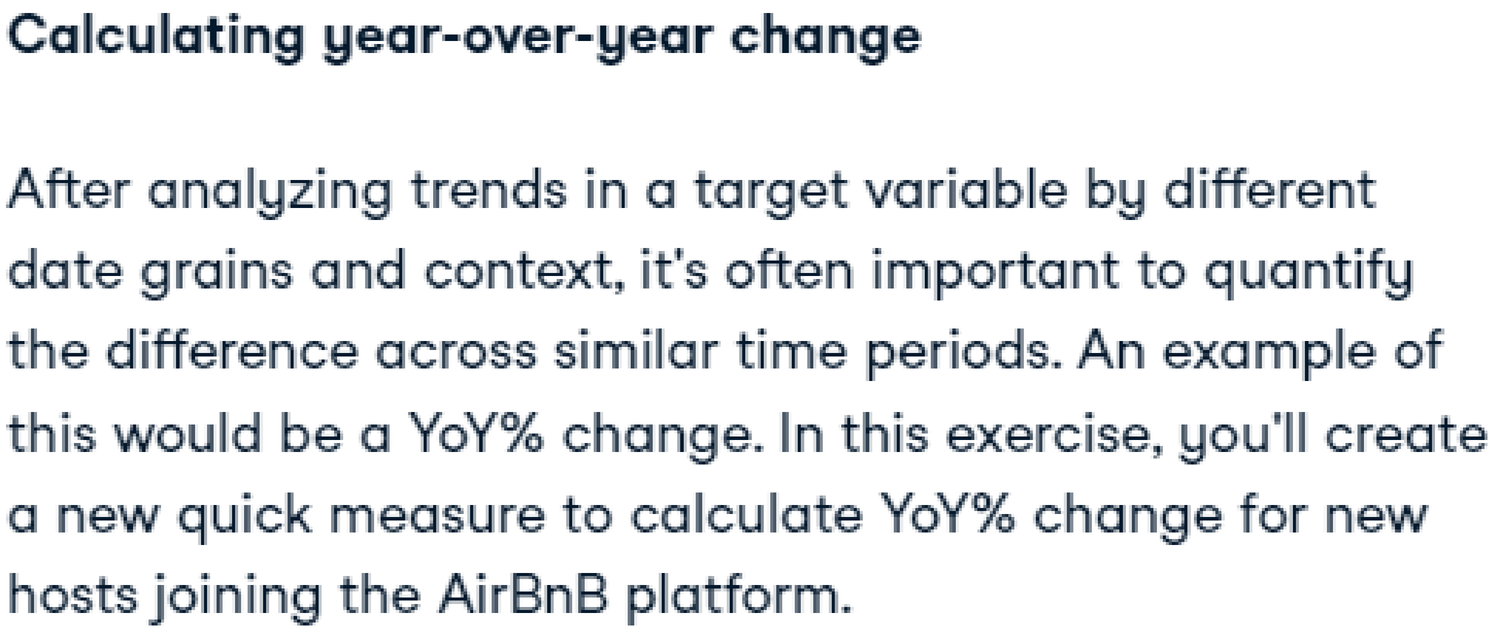
○ 7 - Saturday

○ 1 - Sunday

○ 2 - Monday

○ 5 - Thursday

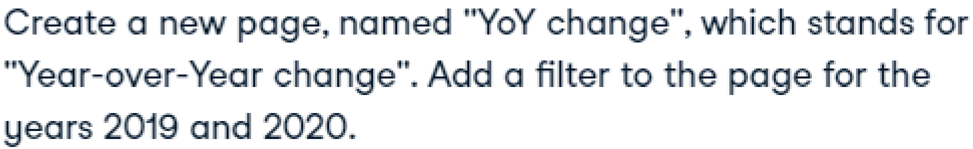
"Complete Exercise"



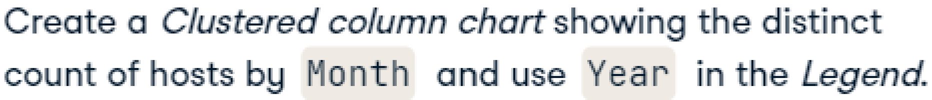
Calculating year-over-year change

After analyzing trends in a target variable by different date grains and context, it's often important to quantify the difference across similar time periods. An example of this would be a YoY% change. In this exercise, you'll create a new quick measure to calculate YoY% change for new hosts joining the AirBnB platform.

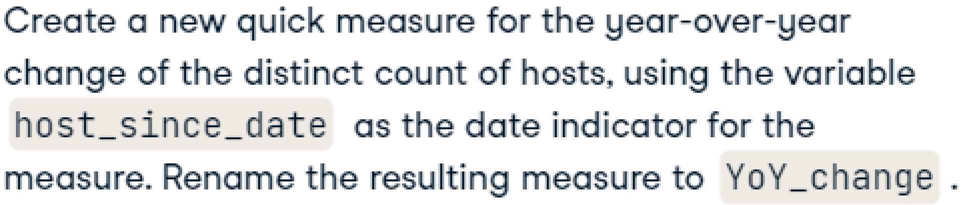
“Instructions"



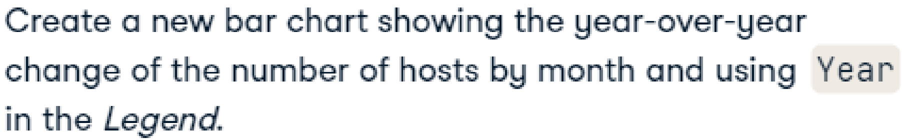
Create a new page, named "YoY change", which stands for "Year-over-Year change". Add a filter to the page for the years 2019 and 2020.



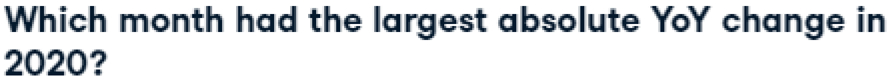
Create a *Clustered column chart* showing the distinct count of hosts by `Month` and use `Year` in the *Legend*.



Create a new quick measure for the year-over-year change of the distinct count of hosts, using the variable `host_since_date` as the date indicator for the measure. Rename the resulting measure to `YoY_change` .



Create a new bar chart showing the year-over-year change of the number of hosts by month and using `Year` in the *Legend*.

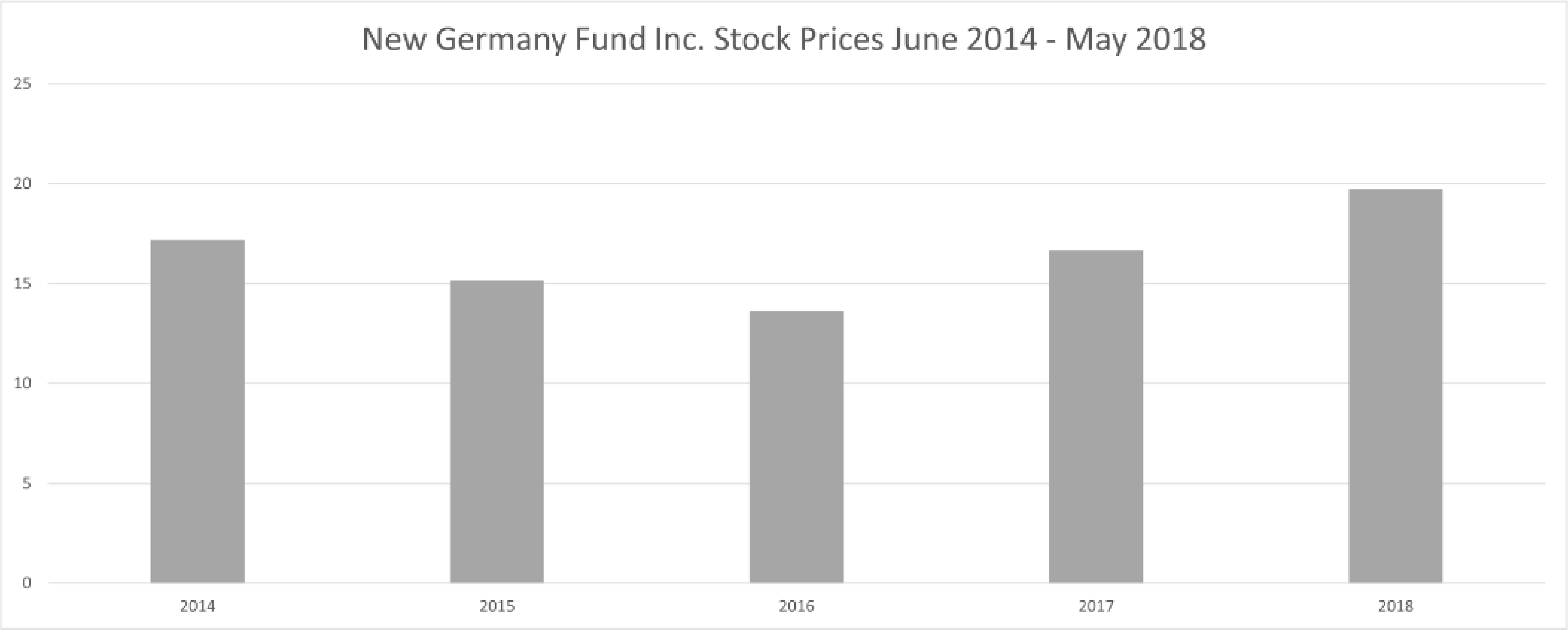


Which month had the largest absolute YoY change in 2020?

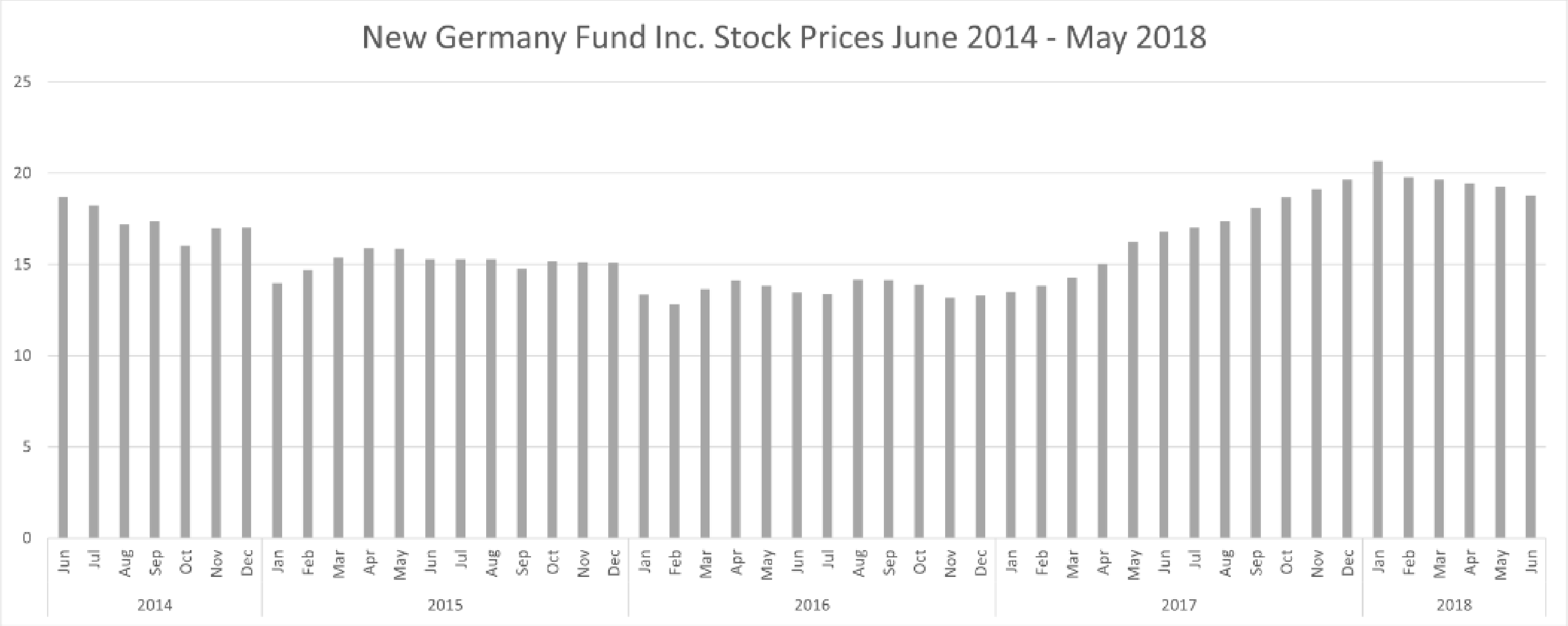
Analyzing Time Series in Power BI

LEARNING VOYAGE

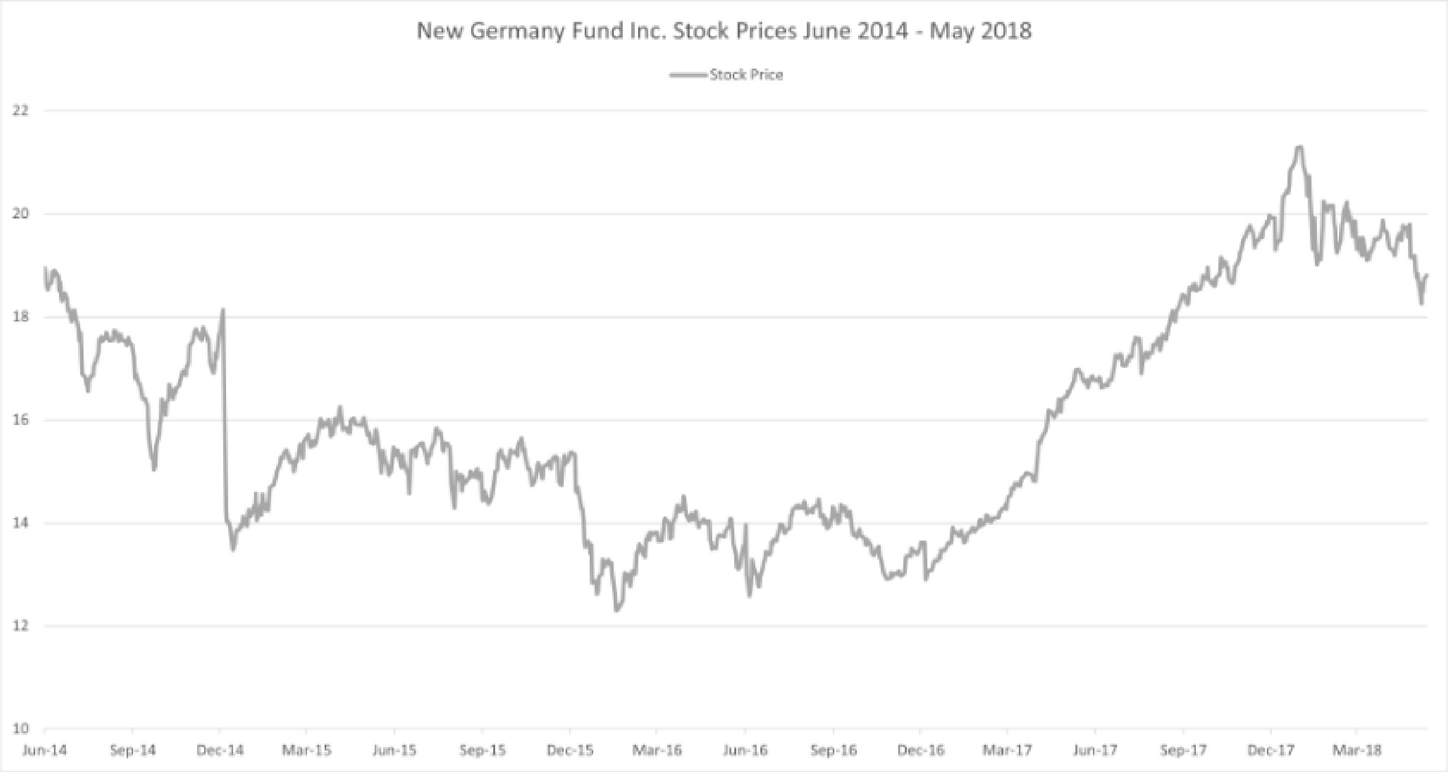
Visualizing a time series



Visualizing a time series

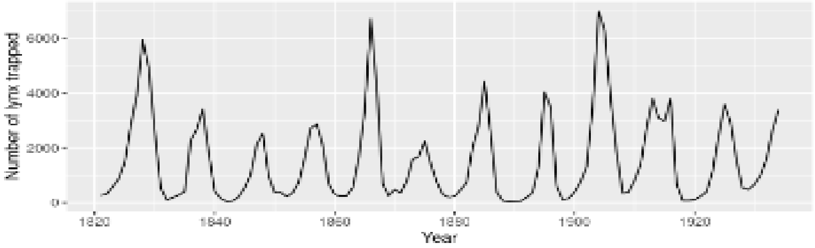


What is a run chart?



Patterns in a time series

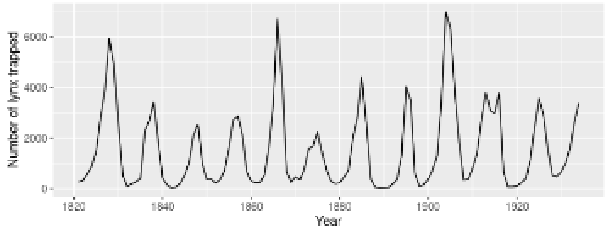
Cyclical



- Rise and fall patterns
- No fixed time period
- Pattern typically longer than a year in length
- Less predictable

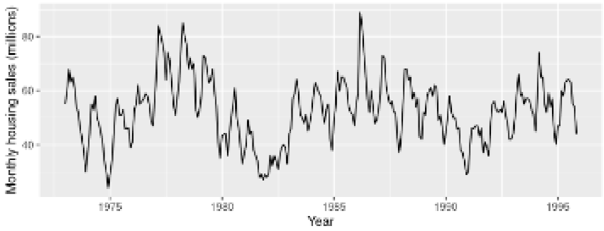
Patterns in a time series

Cyclical



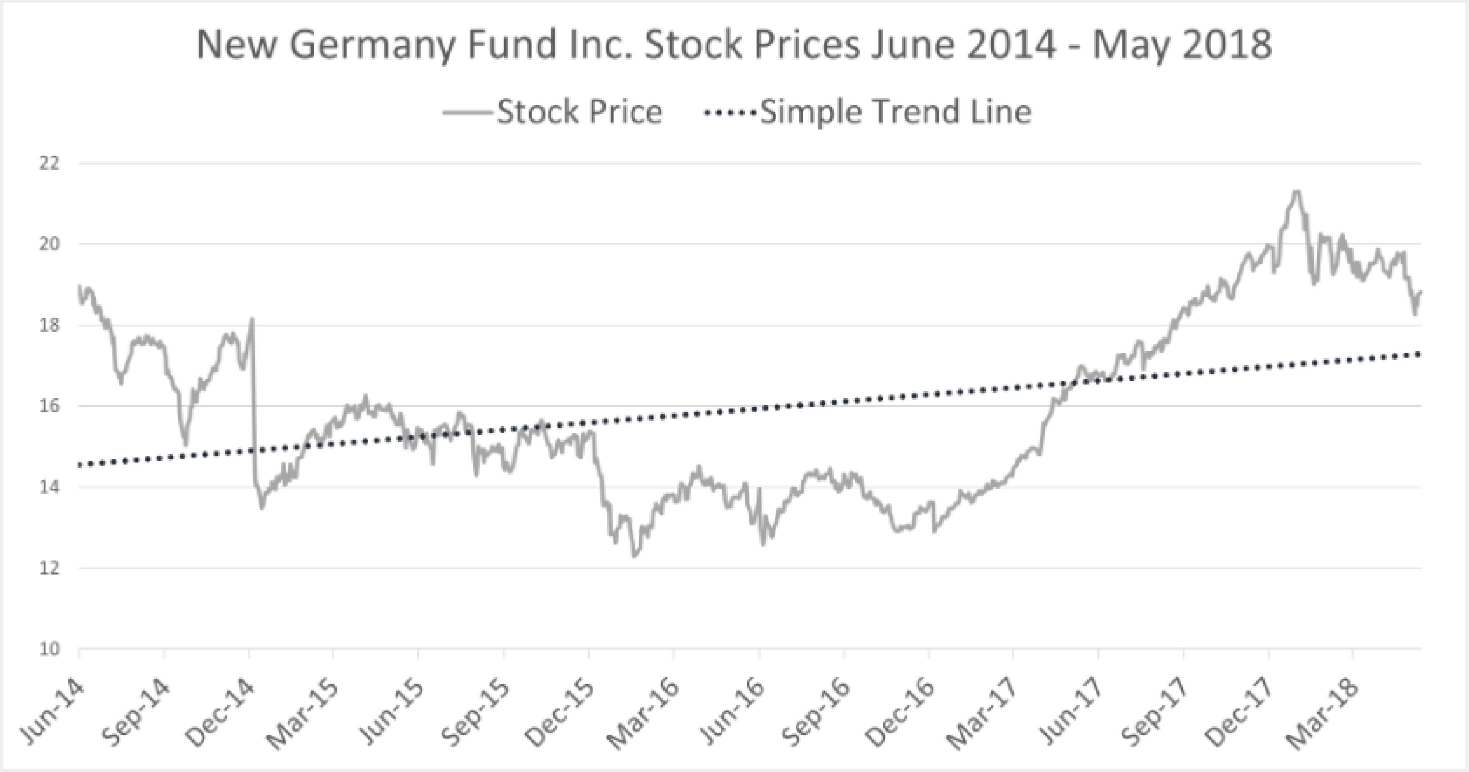
- Rise and fall patterns
- No fixed time period
- Pattern typically longer than a year in length
- Less predictable

Seasonality

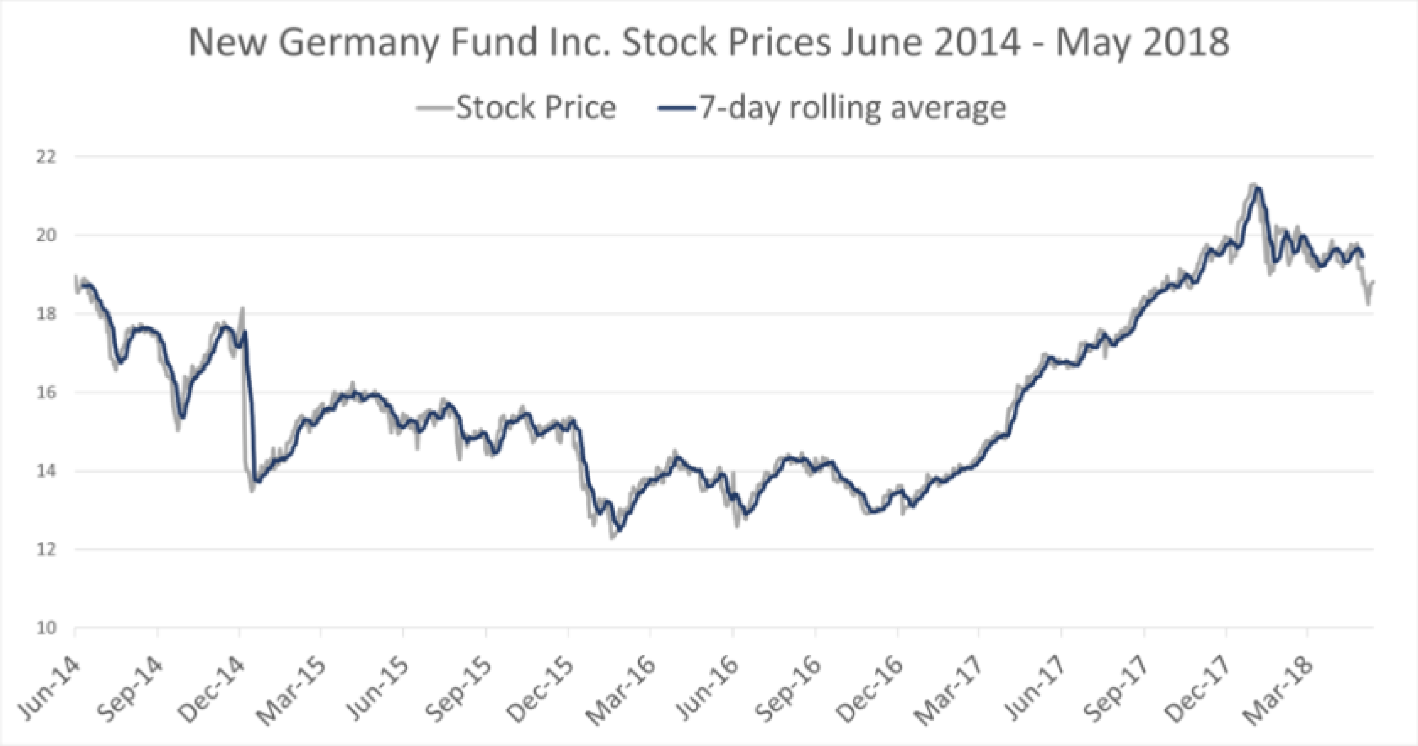


- Pattern is influenced by the season (ex. holiday spending)
- Fixed time periods
- Pattern typically occurs over less than a year

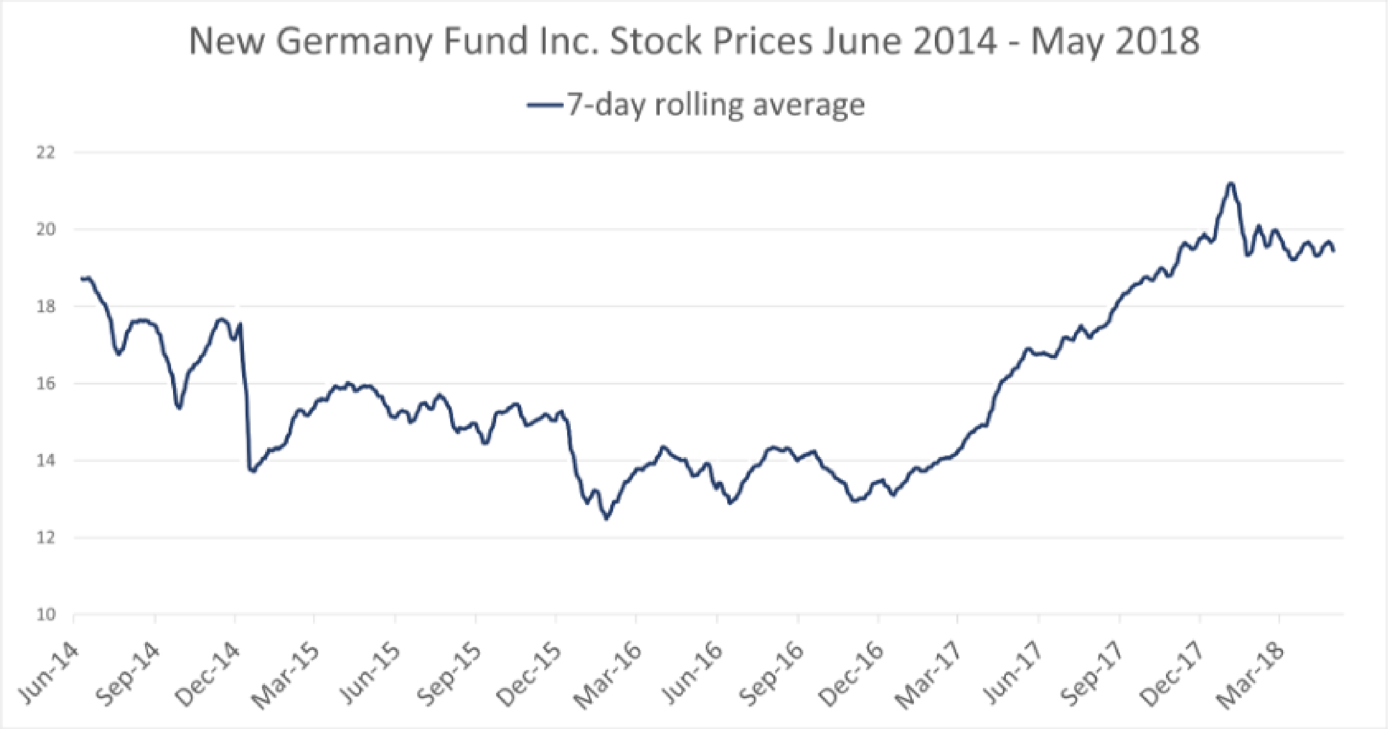
Evaluating the trend in the time series



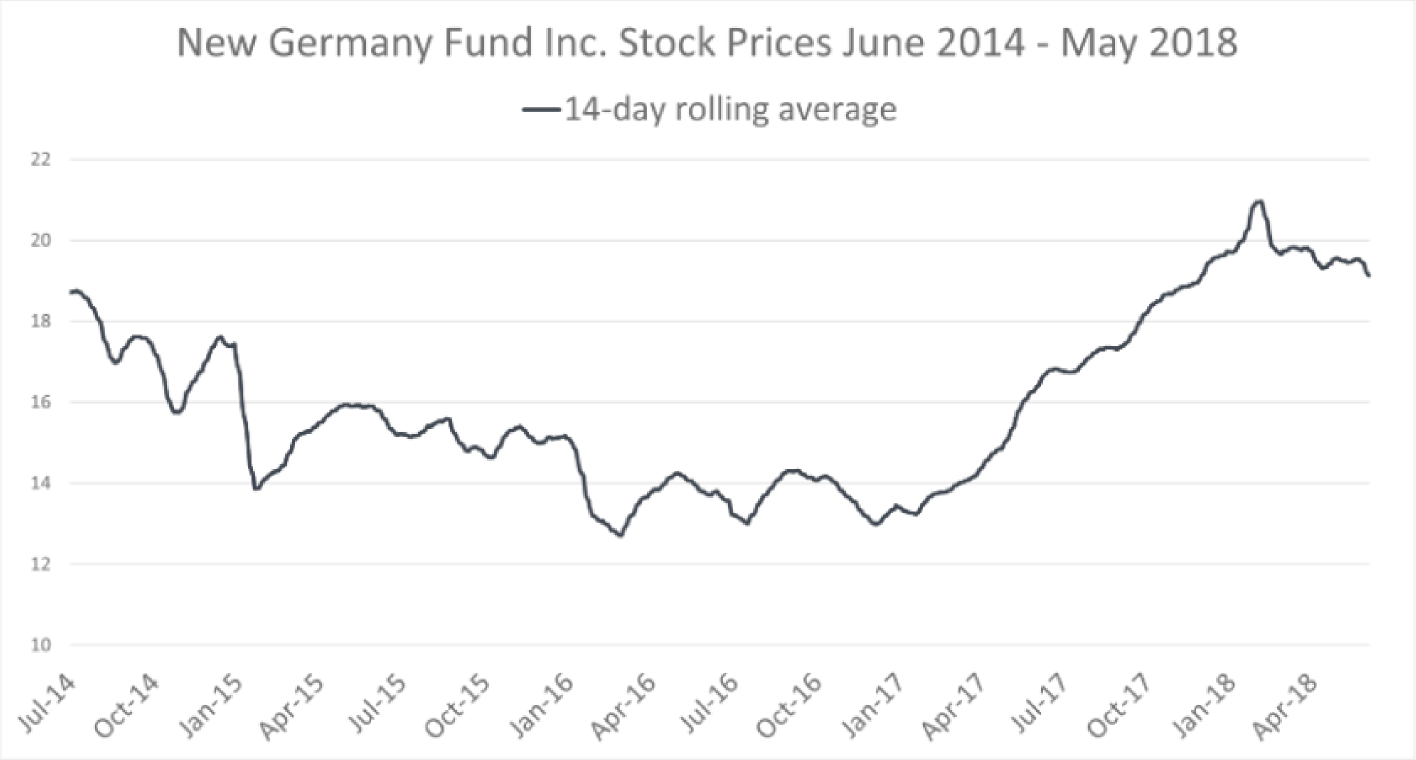
Trends with rolling averages



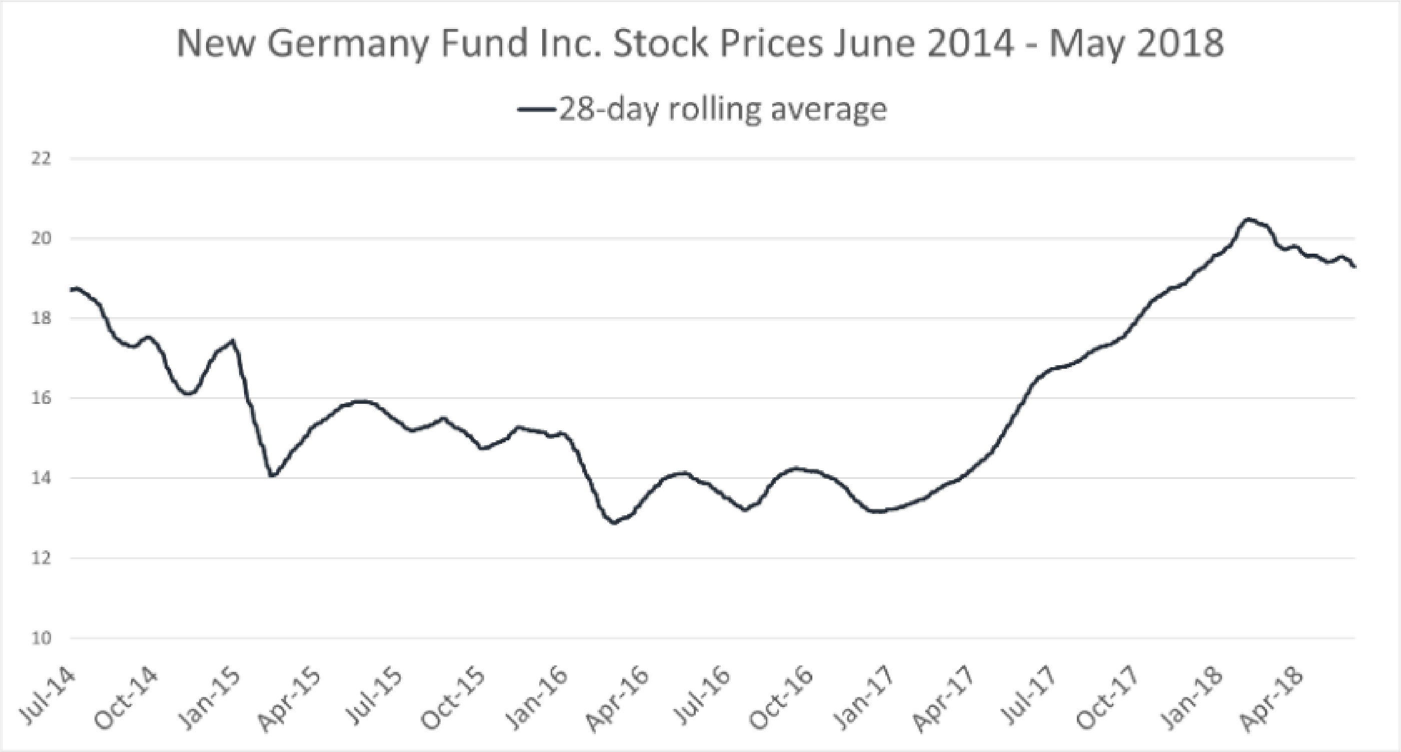
Smoothing with rolling averages



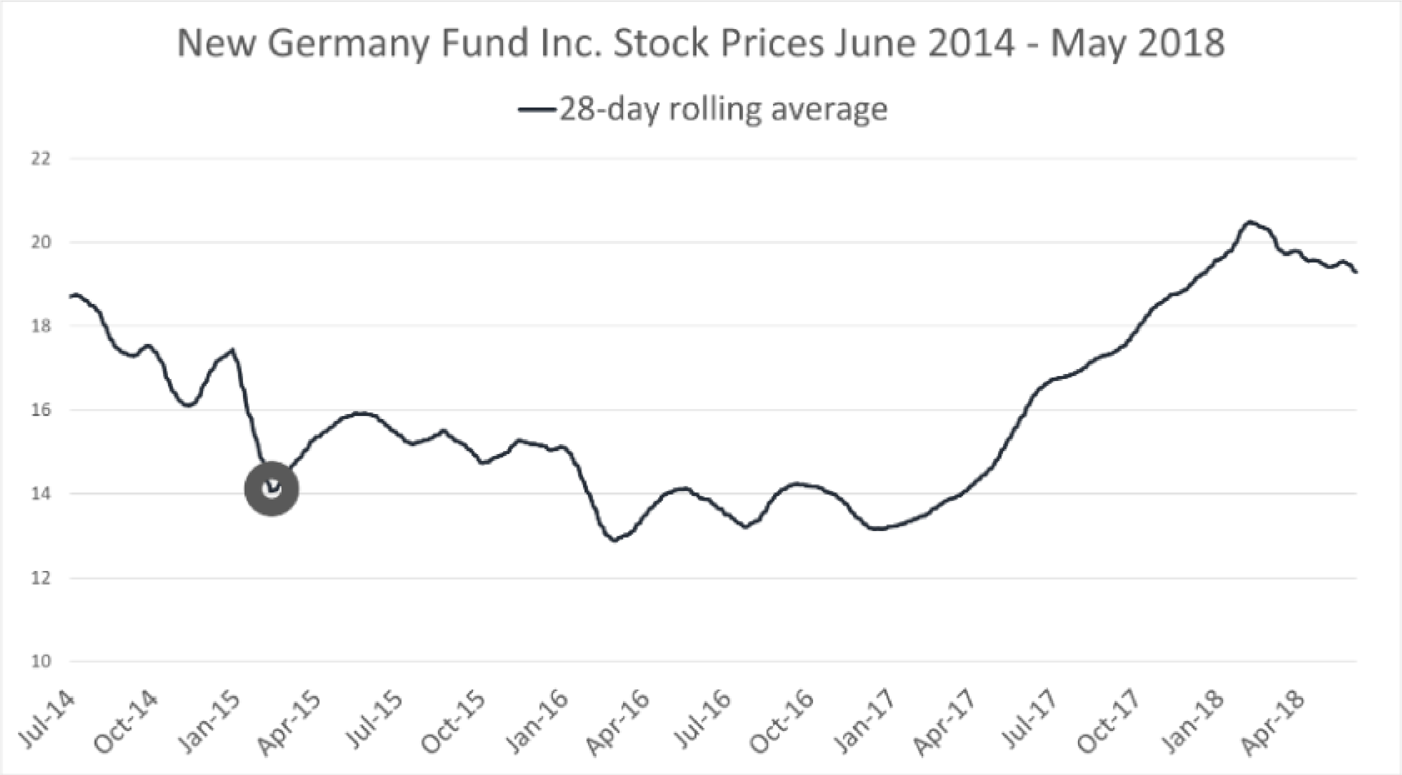
Smoothing with rolling averages



Smoothing with rolling averages



Anomalies in a time series



Let's practice!

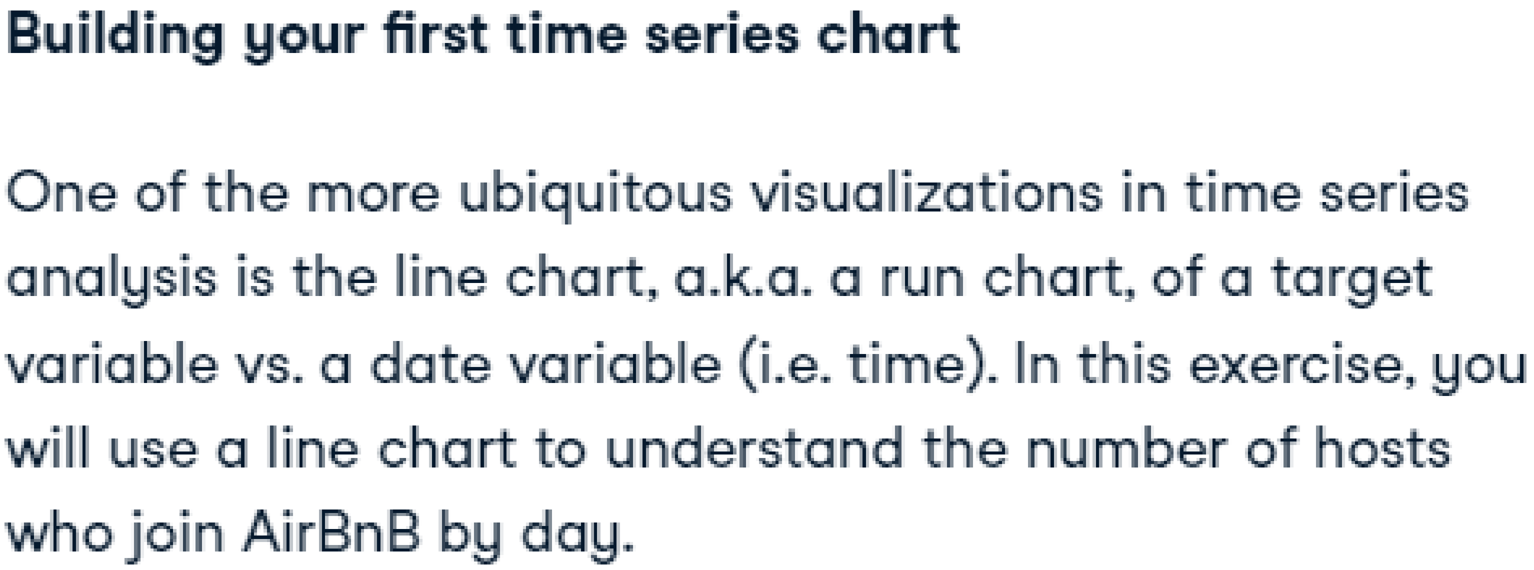
"Cyclical vs. Seasonal"

Being able to read time series line charts will allow greater understanding of the patterns observed within the data. Distinguishing between cyclical and seasonal patterns will help direct further investigation into what is potentially influencing those patterns.

DEMO

Analyzing Glassdoor reviews over time

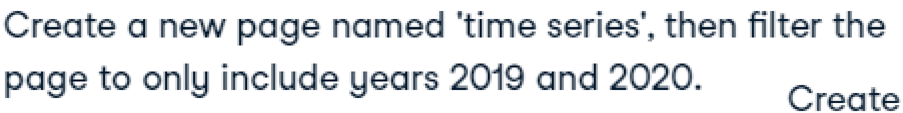
"Complete Exercise"



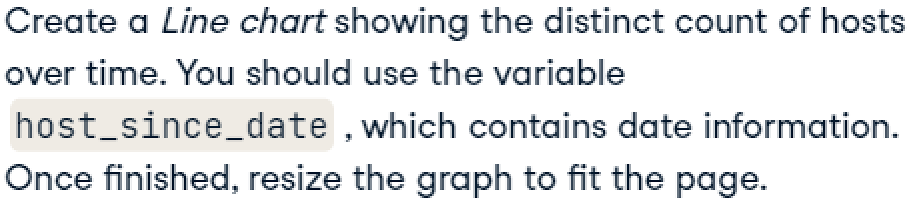
Building your first time series chart

One of the more ubiquitous visualizations in time series analysis is the line chart, a.k.a. a run chart, of a target variable vs. a date variable (i.e. time). In this exercise, you will use a line chart to understand the number of hosts who join AirBnB by day.

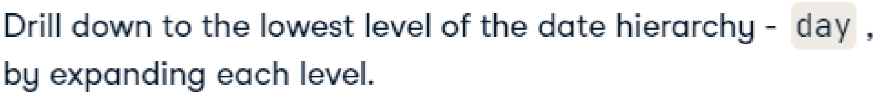
“Instructions"



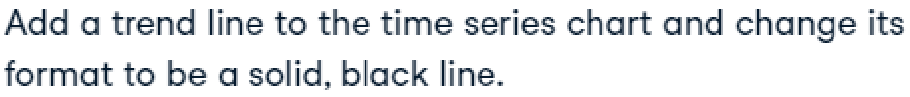
Create a new page named 'time series', then filter the page to only include years 2019 and 2020.



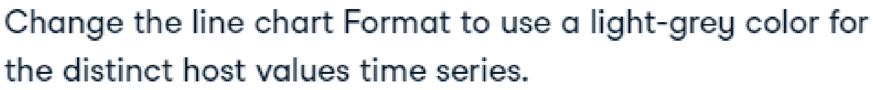
Create a *Line chart* showing the distinct count of hosts over time. You should use the variable `host_since_date` , which contains date information. Once finished, resize the graph to fit the page.



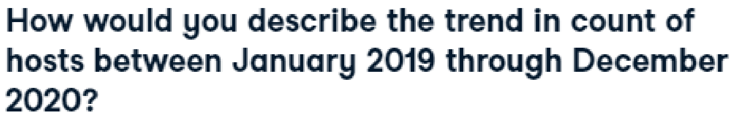
Drill down to the lowest level of the date hierarchy - `day` , by expanding each level.



Add a trend line to the time series chart and change its format to be a solid, black line.

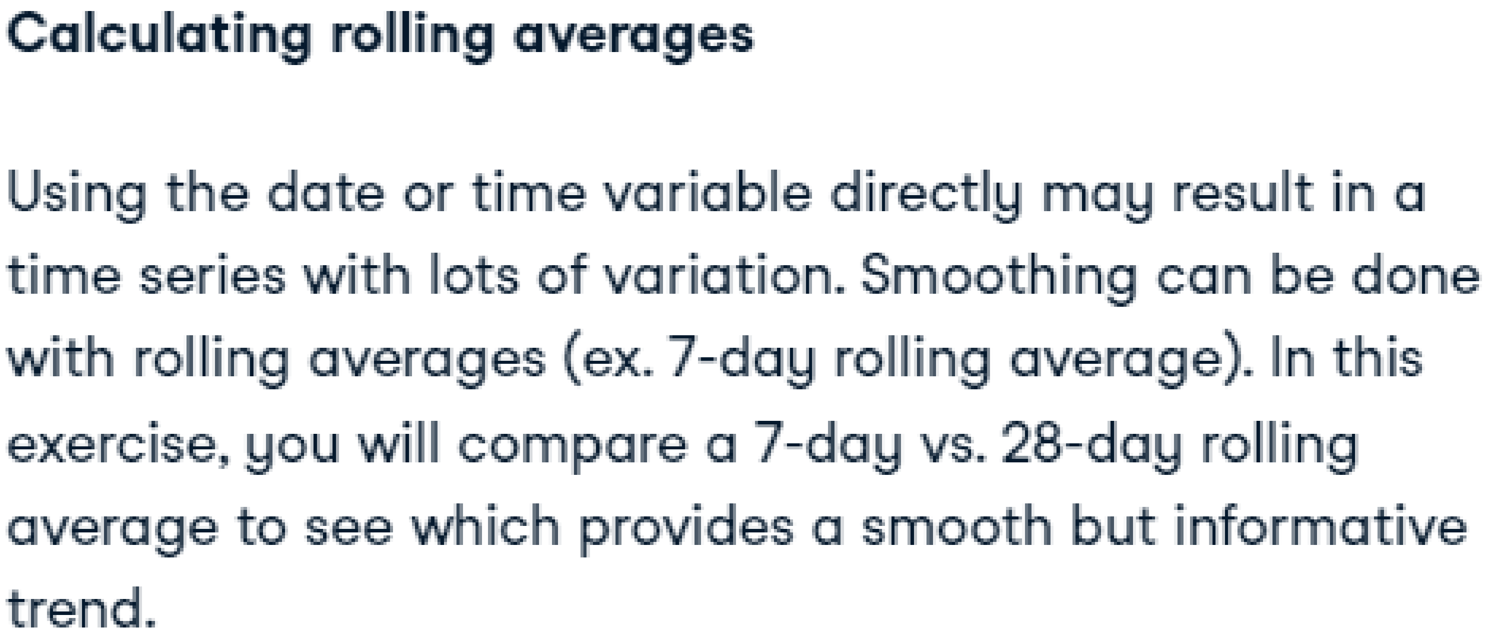


Change the line chart Format to use a light-grey color for the distinct host values time series.



How would you describe the trend in count of hosts between January 2019 through December 2020?

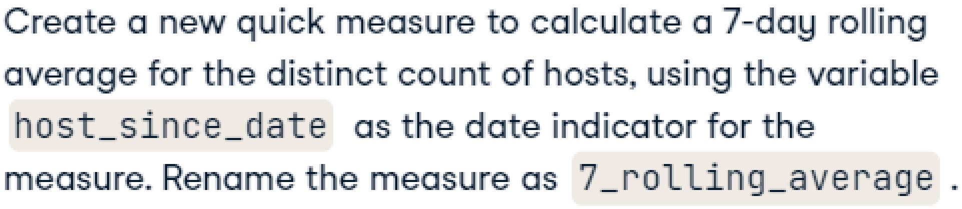
"Complete Exercise"



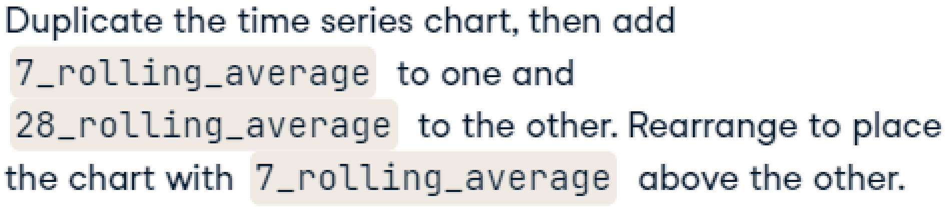
Calculating rolling averages

Using the date or time variable directly may result in a time series with lots of variation. Smoothing can be done with rolling averages (ex. 7-day rolling average). In this exercise, you will compare a 7-day vs. 28-day rolling average to see which provides a smooth but informative trend.

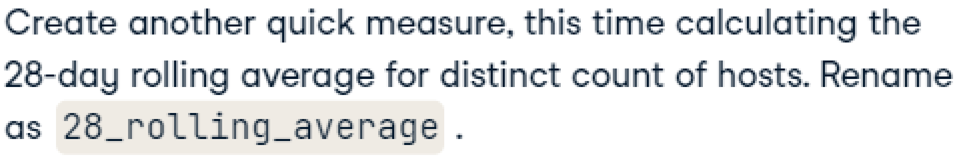
“Instructions"



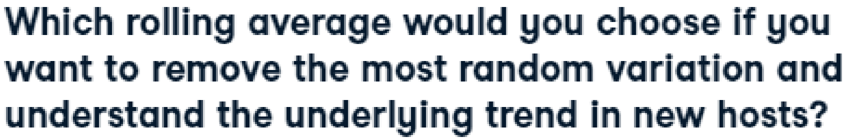
Create a new quick measure to calculate a 7-day rolling average for the distinct count of hosts, using the variable `host_since_date` as the date indicator for the measure. Rename the measure as `7_rolling_average` .



Duplicate the time series chart, then add `7_rolling_average` to one and `28_rolling_average` to the other. Rearrange to place the chart with `7_rolling_average` above the other.

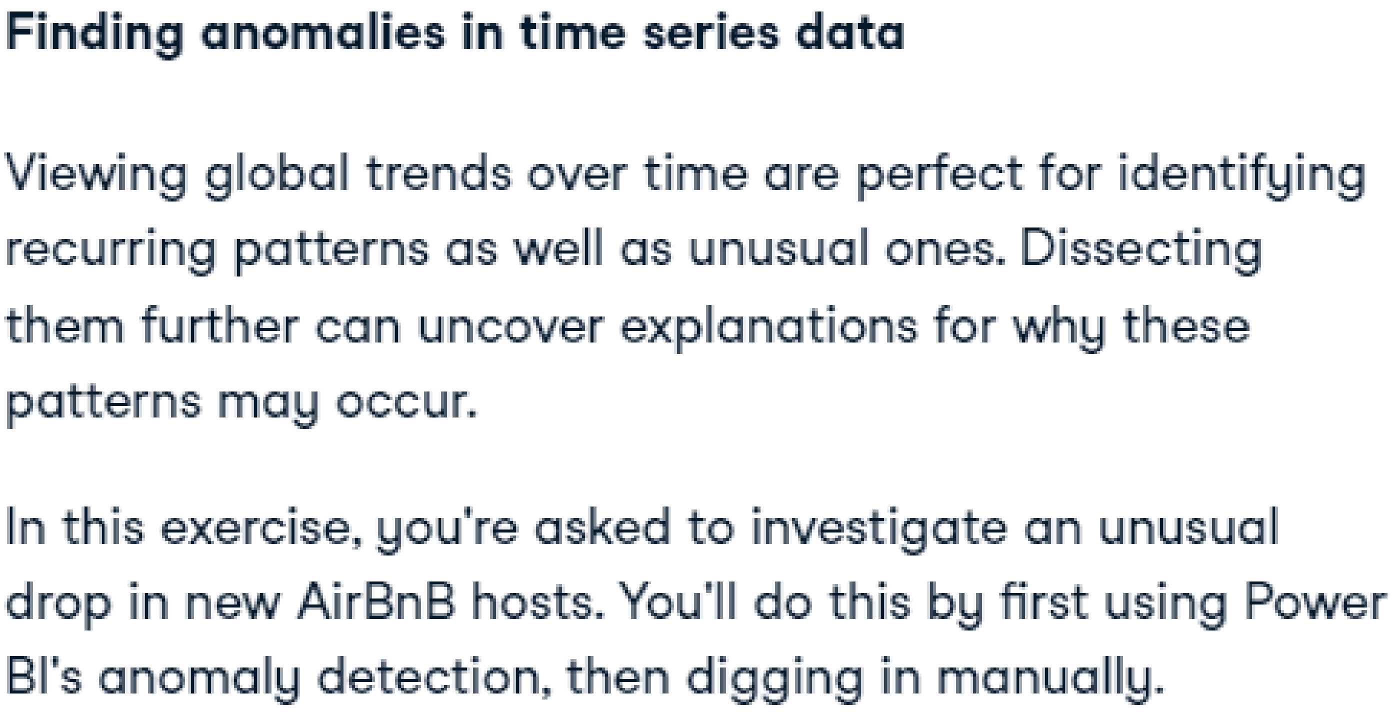


Create another quick measure, this time calculating the 28-day rolling average for distinct count of hosts. Rename as `28_rolling_average` .



Which rolling average would you choose if you want to remove the most random variation and understand the underlying trend in new hosts?

"Complete Exercise"

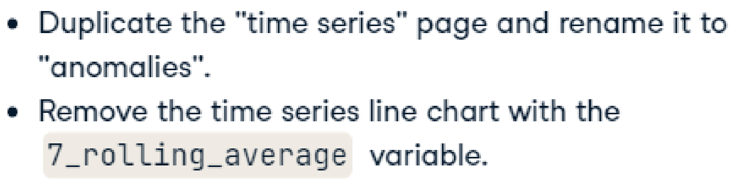


Finding anomalies in time series data

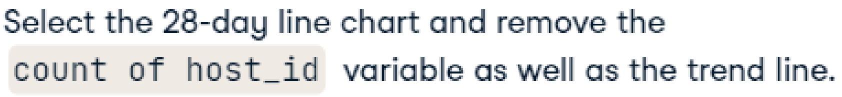
Viewing global trends over time are perfect for identifying recurring patterns as well as unusual ones. Dissecting them further can uncover explanations for why these patterns may occur.

In this exercise, you're asked to investigate an unusual drop in new AirBnB hosts. You'll do this by first using Power BI's anomaly detection, then digging in manually.

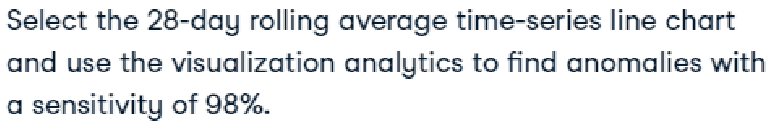
“Instructions"



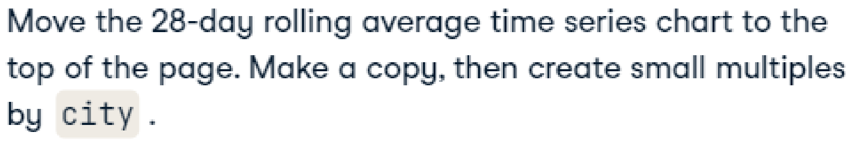
- Duplicate the "time series" page and rename it to "anomalies".
- Remove the time series line chart with the `7_rolling_average` variable.



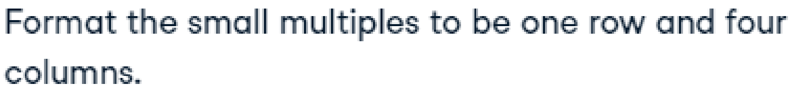
Select the 28-day line chart and remove the `count of host_id` variable as well as the trend line.



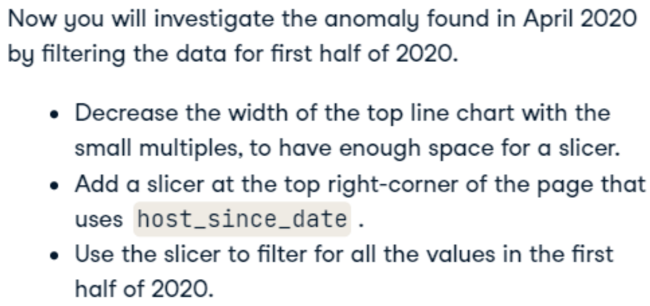
Select the 28-day rolling average time-series line chart and use the visualization analytics to find anomalies with a sensitivity of 98%.



Move the 28-day rolling average time series chart to the top of the page. Make a copy, then create small multiples by `city` .

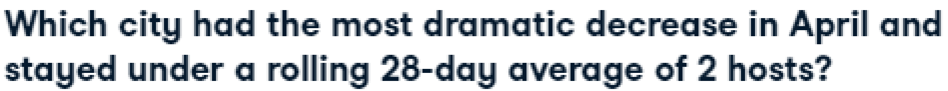


Format the small multiples to be one row and four columns.



Now you will investigate the anomaly found in April 2020 by filtering the data for first half of 2020.

- Decrease the width of the top line chart with the small multiples, to have enough space for a slicer.
- Add a slicer at the top right-corner of the page that uses `host_since_date` .
- Use the slicer to filter for all the values in the first half of 2020.



Which city had the most dramatic decrease in April and stayed under a rolling 28-day average of 2 hosts?

Decomposition Trees

What is a decomposition tree?

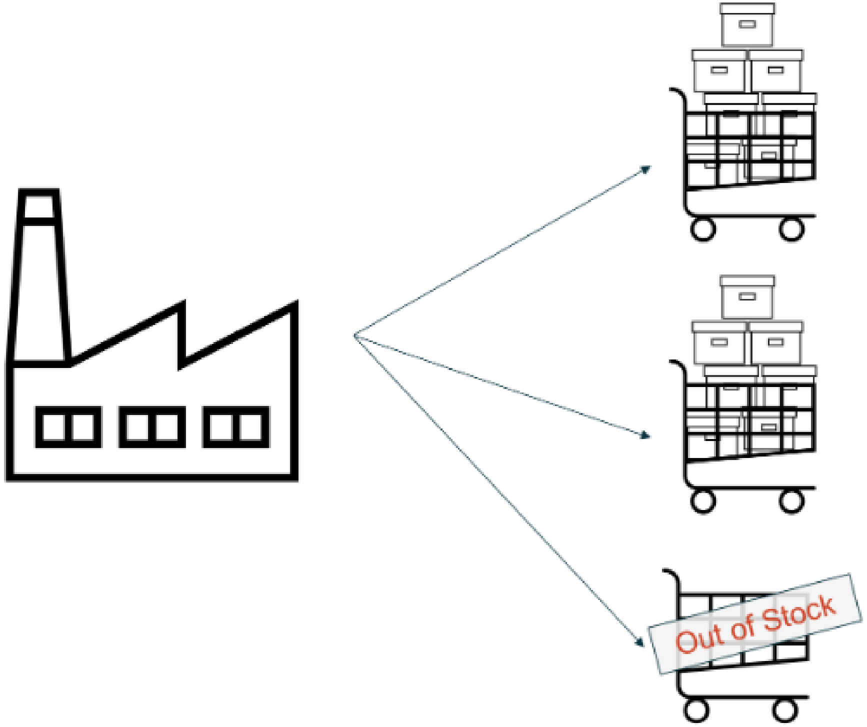
Definition:

A method for breaking down a target variable by multiple dimensions or variables to determine resulting influence.

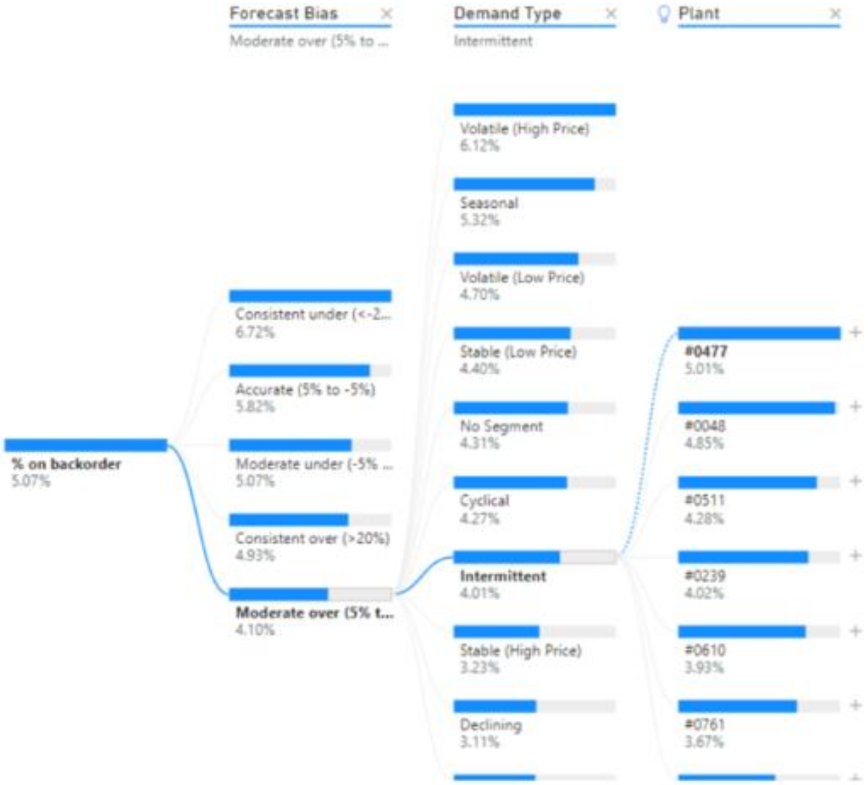
Use Cases:

- Ad hoc exploration
- Root cause analysis
- Identifying influential variables

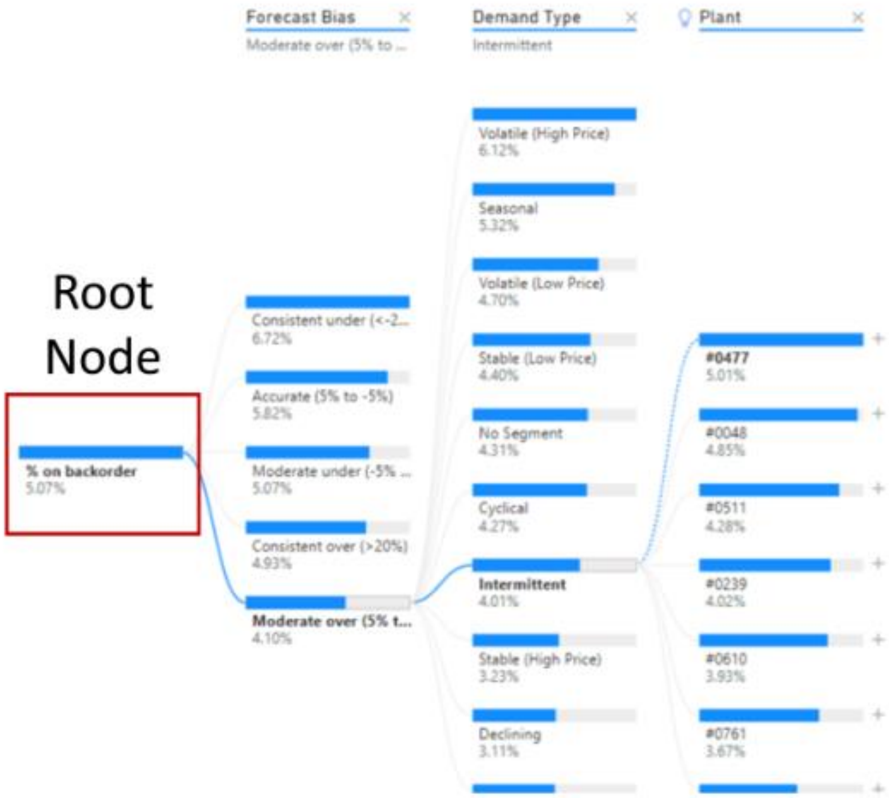
Structure of a decomposition tree



Structure of a decomposition tree

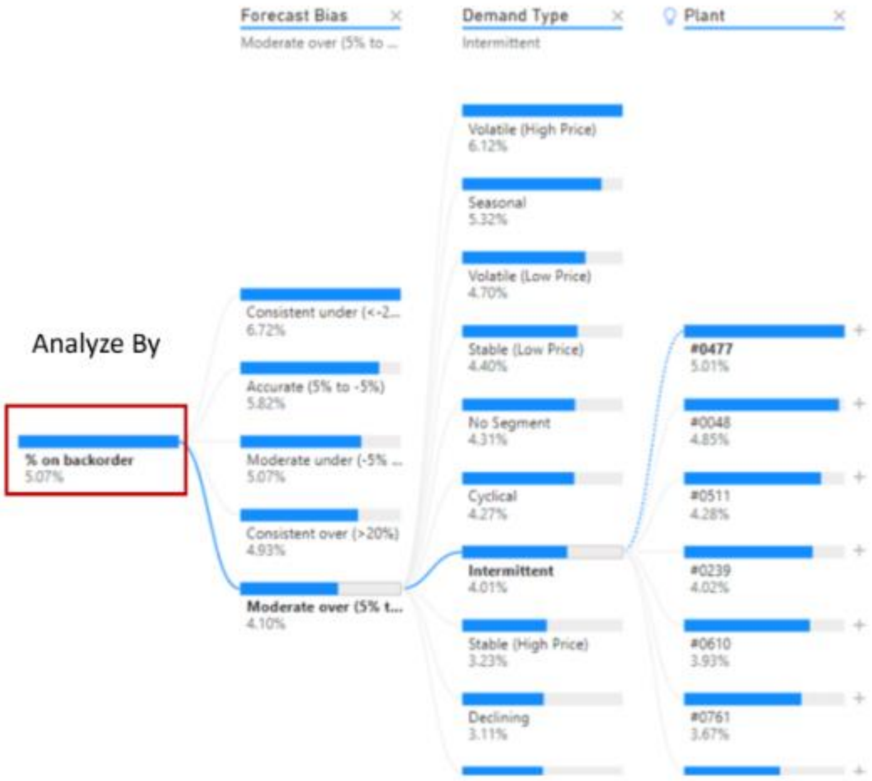


Structure of a decomposition tree

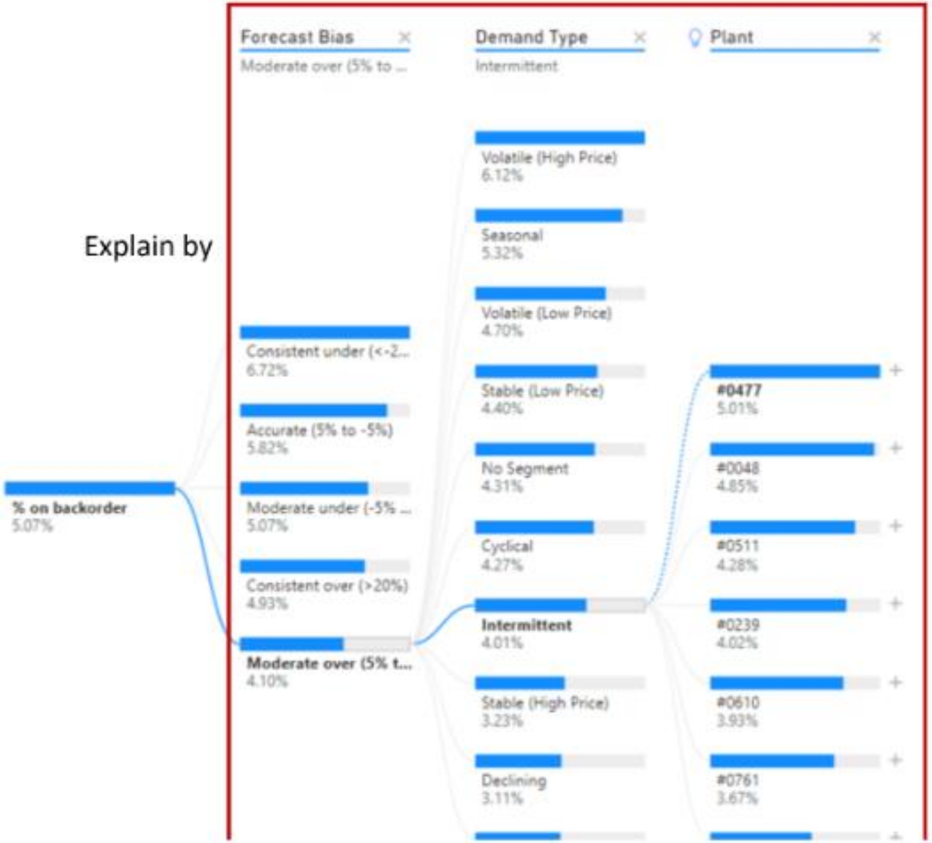


Root Node

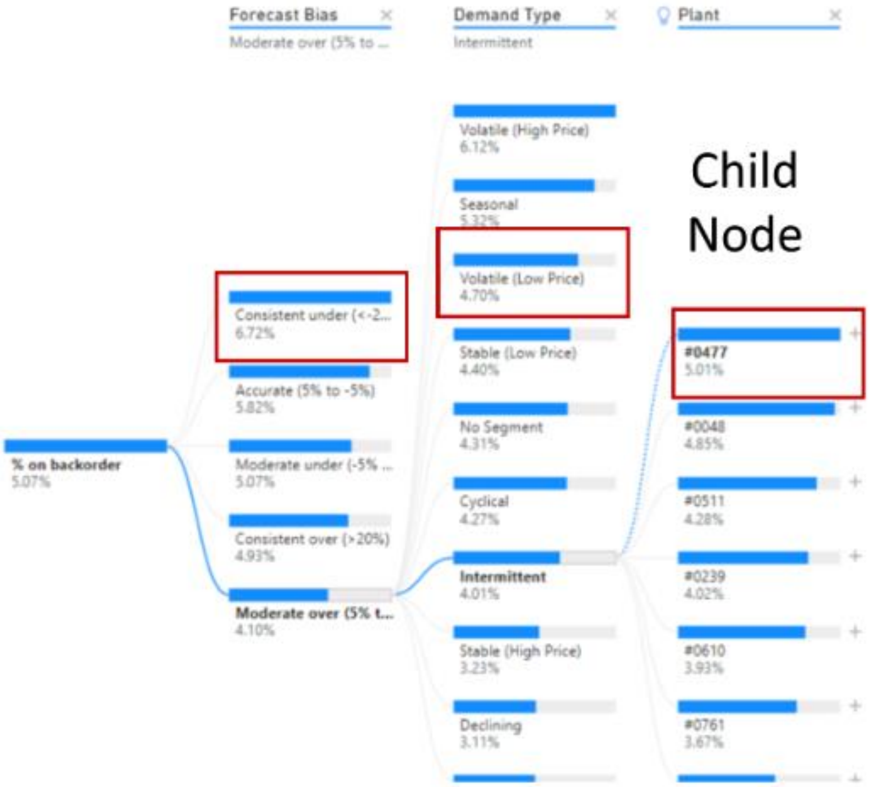
Structure of a decomposition tree



Structure of a decomposition tree

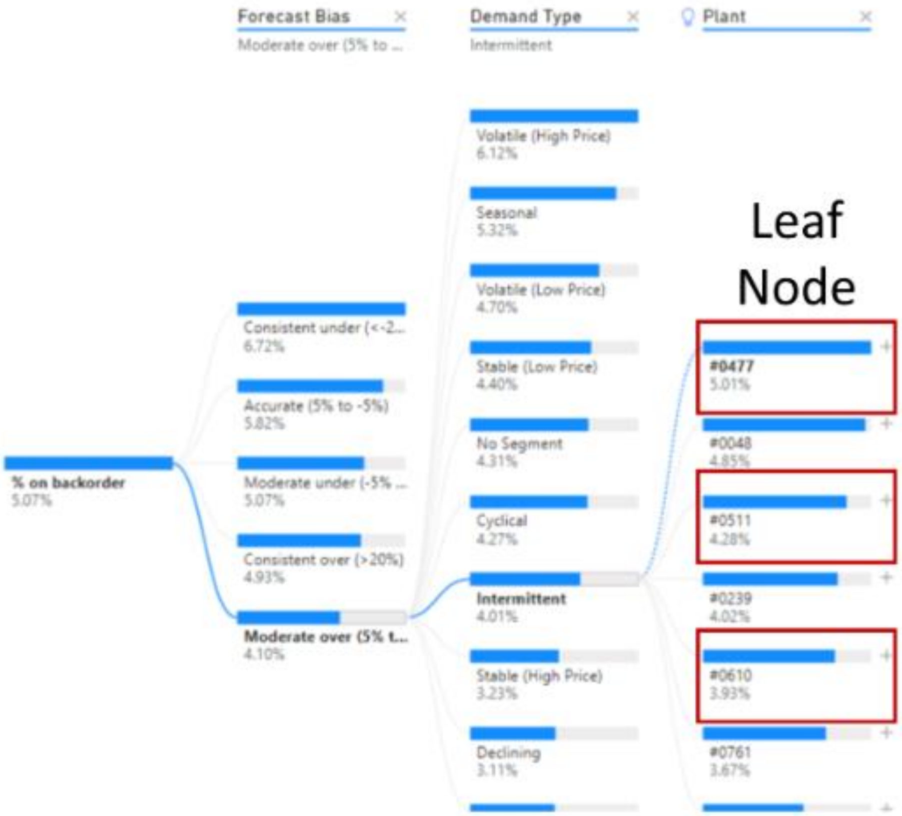

Explain by

Structure of a decomposition tree



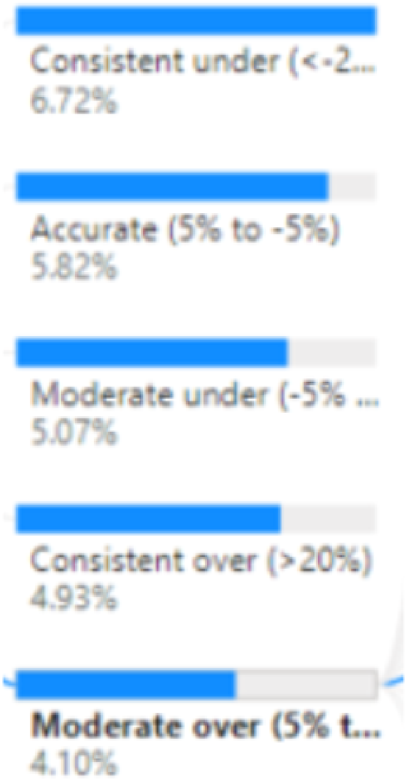
Child Node

Structure of a decomposition tree

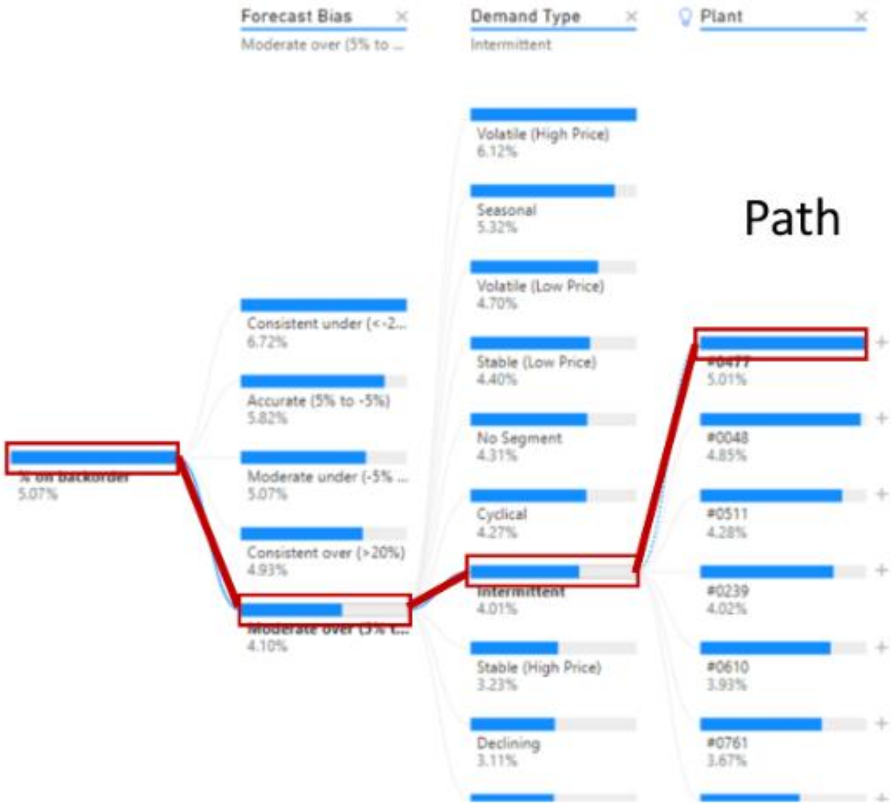


Leaf Node

Structure of a decomposition tree

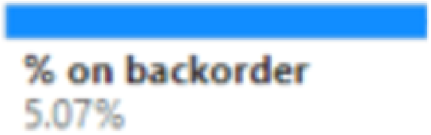


Structure of a decomposition tree



Path

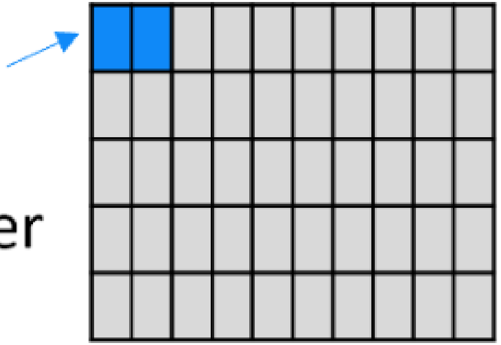
Reading a decomposition tree



% on backorder

5.07%

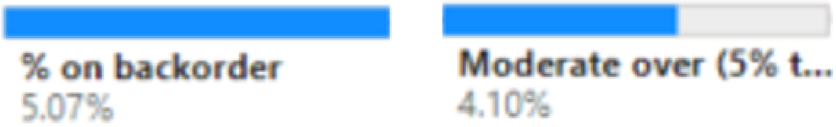
ALL ORDERS



5.07%
on
backorder

Reading a decomposition tree

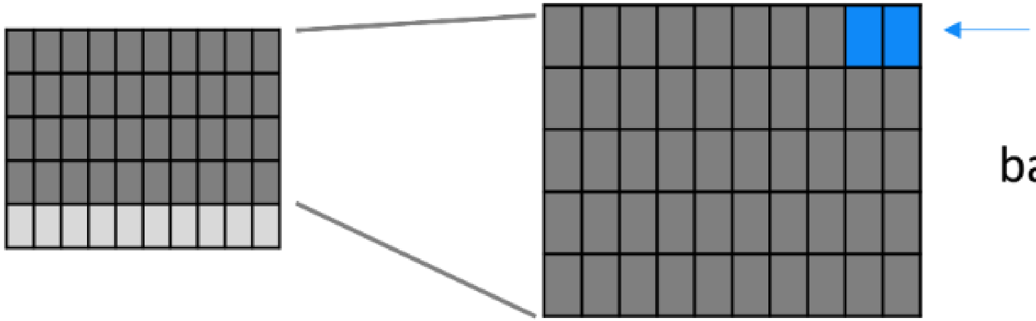
Forecast Bias



% on backorder
5.07%

Moderate over (5% t...
4.10%

SUBSET OF ORDERS

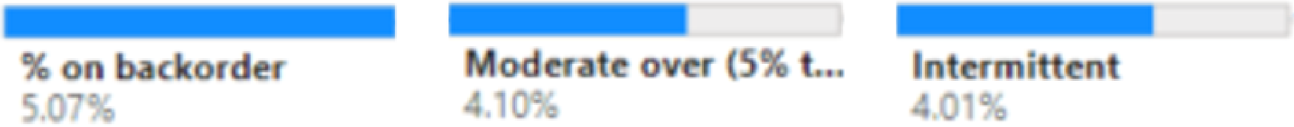


4.10%
on
backorder

Reading a decomposition tree

Forecast Bias

Demand Type

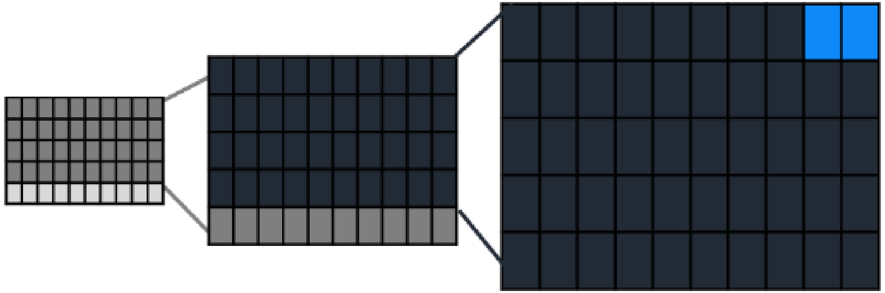


% on backorder
5.07%

Moderate over (5% t...
4.10%

Intermittent
4.01%

ORDERS w/ Moderate Forecast Bias AND Intermittent Demand



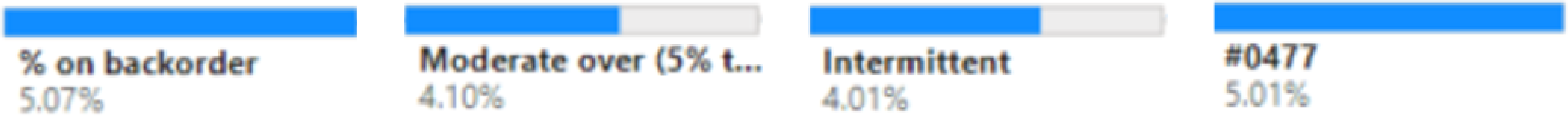
4.01% on backorder

Reading a decomposition tree

Forecast Bias

Demand Type

Plant



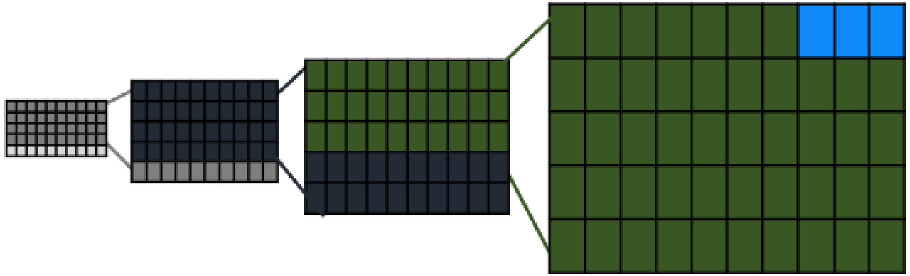
% on backorder
5.07%

Moderate over (5% t...
4.10%

Intermittent
4.01%

#0477
5.01%

ORDERS w/ Moderate
Forecast Bias AND
Intermittent Demand AND
Plant #0477



5.01%
on
backorder

Statements about decomposition trees

Which of the following statements about the use of a decomposition tree is NOT true?

- ○ They are great for understanding what causes a metric to decline.
- ○ Decomposition trees are fantastic for visualizing the distribution of a variable.
- ○ They can be useful for ad hoc exploration of a variable.

DEMO

Using decomposition trees in Power BI

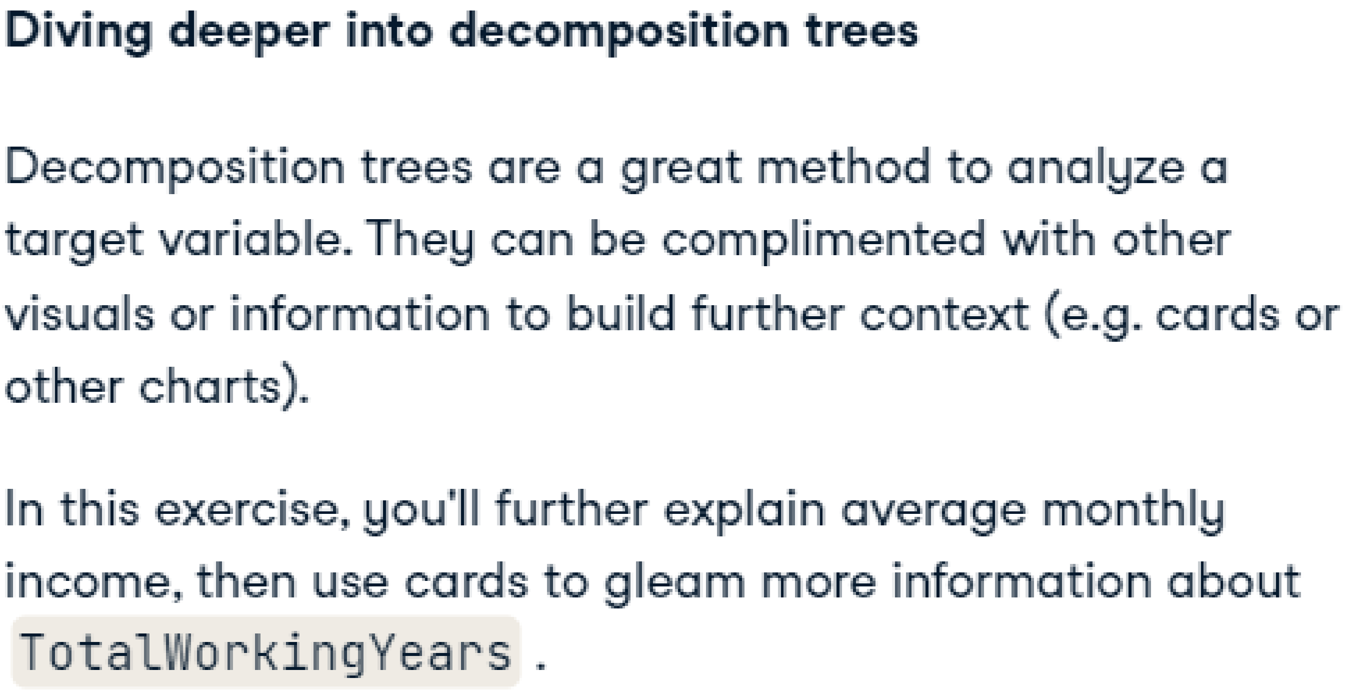
Getting started with decomposition trees

Let's revisit the Glassdoor dataset. The past exercises uncovered relationships between MonthlyIncome and other variables in the dataset including JobRole, JobLevel, and CareerStage. These can be used in a decomposition tree to quicker analysis of how values within each variable influence the value of MonthlyIncome.
In this exercise, you'll do just that to understand how average MonthlyIncome is influenced by different values of several "explain by" variables.

Instructions

1. Close all reports, then open 3_1_decomposition.pbix.
2. On a new page, named "decomposition", add a new decomposition tree analyzing average Monthly Income by JobLevel, JobRole, CareerStage, and Gender.
3. Expand the visual to take up the 80% of the page, then drill into average monthly income by choosing the Explain by variables you think are most influential on the average monthly income.
4. Remove the expanded variables by clicking the x next to each Explain by variable. Use the "AI Splits" feature to drill four times into the "Low value", making sure to choose the child node with the lowest value for each split.
5. What is the order of "Explain by" variables the AI splits expanded to based on the lowest value of average monthly income?

"Complete Exercise"

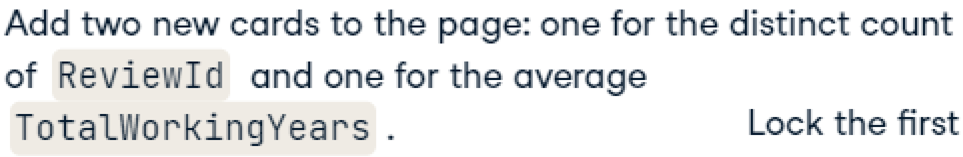


Diving deeper into decomposition trees

Decomposition trees are a great method to analyze a target variable. They can be complimented with other visuals or information to build further context (e.g. cards or other charts).

In this exercise, you'll further explain average monthly income, then use cards to gleam more information about `TotalWorkingYears` .

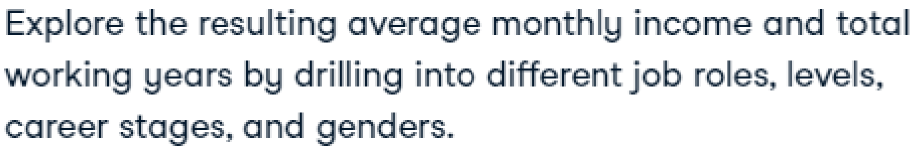
“Instructions"



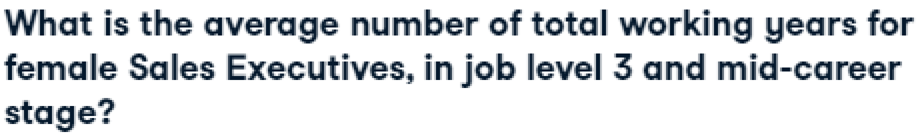
Add two new cards to the page: one for the distinct count of `ReviewId` and one for the average `TotalWorkingYears` .

Lock the first "Explain by" level, `JobRole` , then remove the rest of the expanded *Explain by* variables.

Expand the *Explain by* variables to reveal all of them in the decomposition tree.

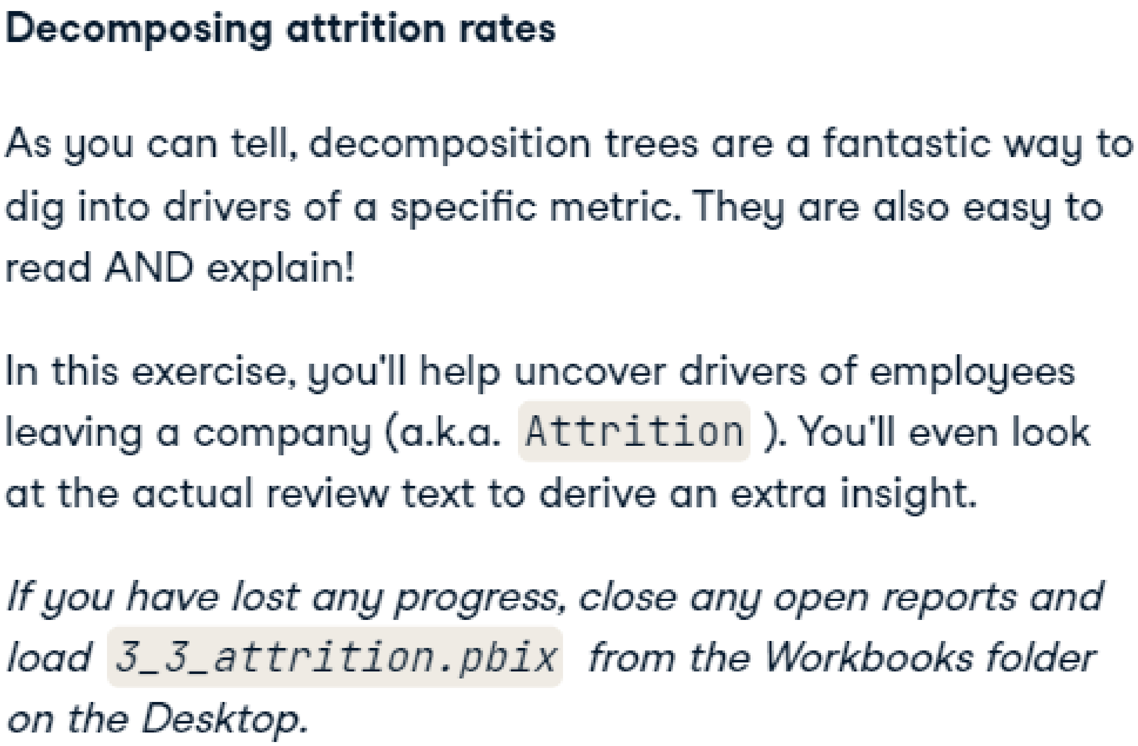


Explore the resulting average monthly income and total working years by drilling into different job roles, levels, career stages, and genders.



What is the average number of total working years for female Sales Executives, in job level 3 and mid-career stage?

"Complete Exercise"



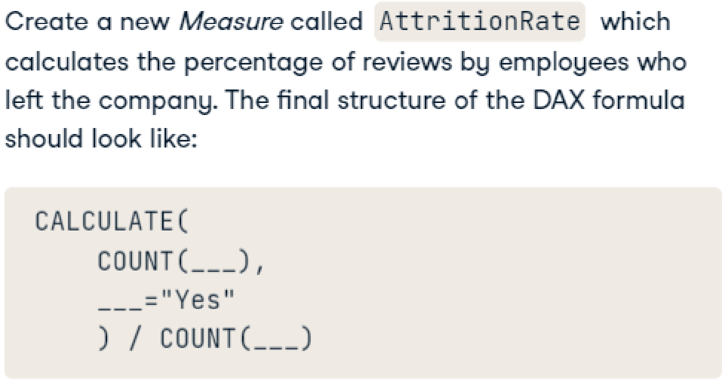
Decomposing attrition rates

As you can tell, decomposition trees are a fantastic way to dig into drivers of a specific metric. They are also easy to read AND explain!

In this exercise, you'll help uncover drivers of employees leaving a company (a.k.a. `Attrition`). You'll even look at the actual review text to derive an extra insight.

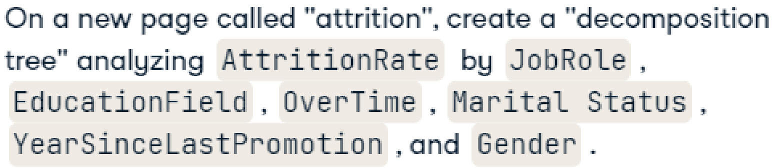
If you have lost any progress, close any open reports and load `3_3_attrition.pbix` *from the Workbooks folder on the Desktop.*

"Instructions"

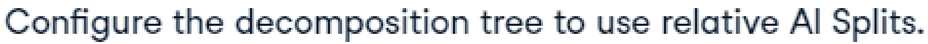


Create a new *Measure* called `AttritionRate` which calculates the percentage of reviews by employees who left the company. The final structure of the DAX formula should look like:

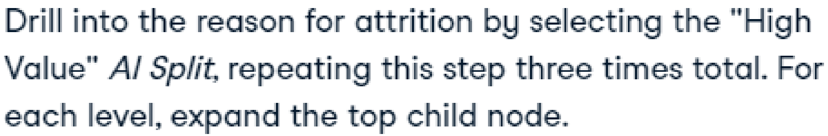
```
CALCULATE(
    COUNT(___),
    ___="Yes"
    ) / COUNT(___)
```



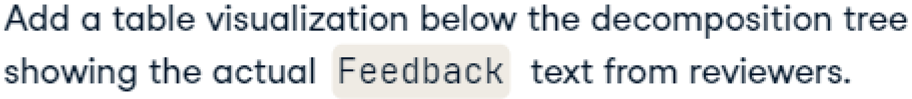
On a new page called "attrition", create a "decomposition tree" analyzing `AttritionRate` by `JobRole`, `EducationField`, `OverTime`, `Marital Status`, `YearSinceLastPromotion`, and `Gender`.



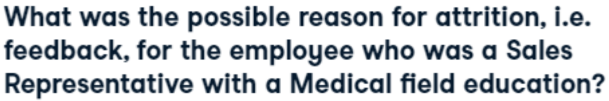
Configure the decomposition tree to use relative AI Splits.



Drill into the reason for attrition by selecting the "High Value" *AI Split*, repeating this step three times total. For each level, expand the top child node.



Add a table visualization below the decomposition tree showing the actual `Feedback` text from reviewers.



What was the possible reason for attrition, i.e. feedback, for the employee who was a Sales Representative with a Medical field education?

Key Influencers

LEARNING
VOYAGE

What are "key influencers"?

Definition

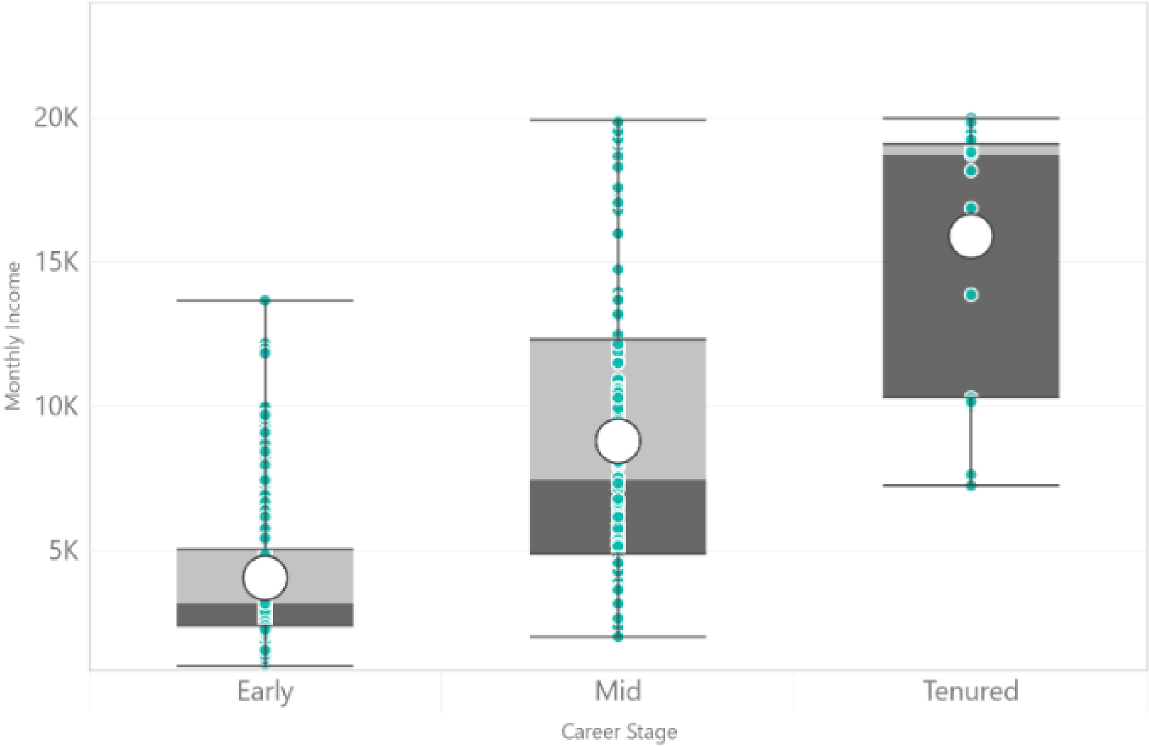
Explanatory variables which cause a significant change in the characteristic of a target variable.

Use Cases

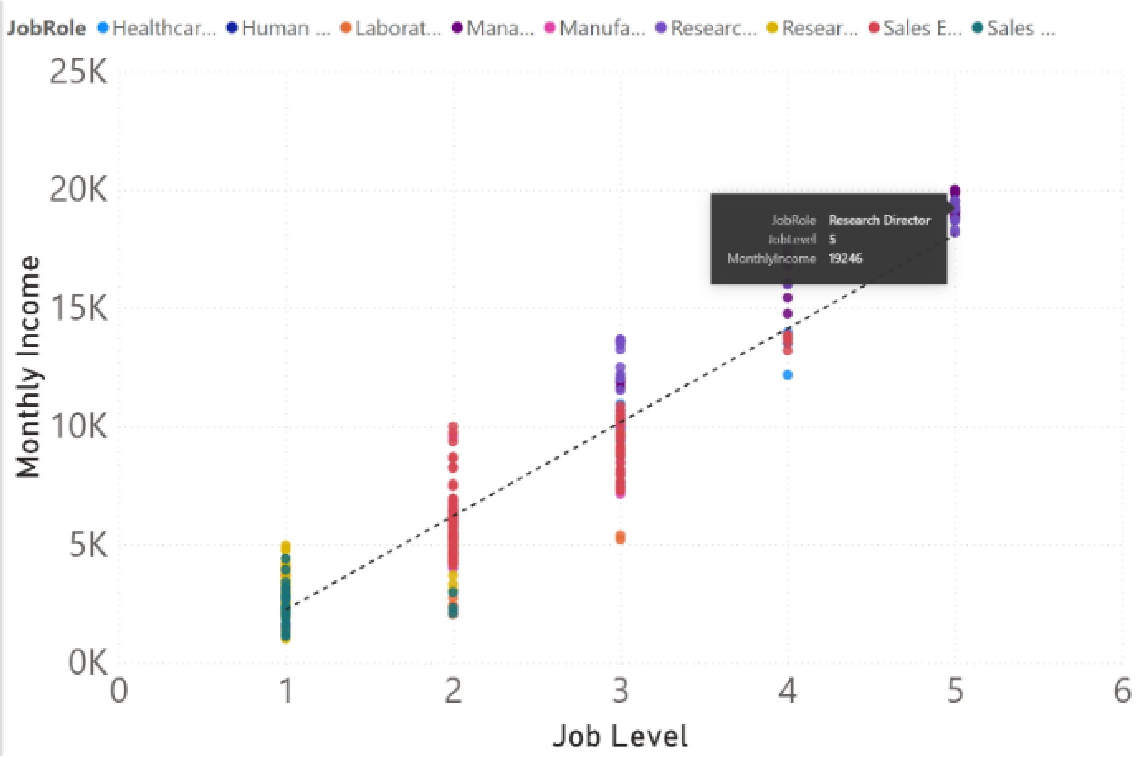
- Determine major drivers of a metric
- Compare the relative importance of variables
- Explore relationships between a target and another variable

[1] https://docs.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-influencers

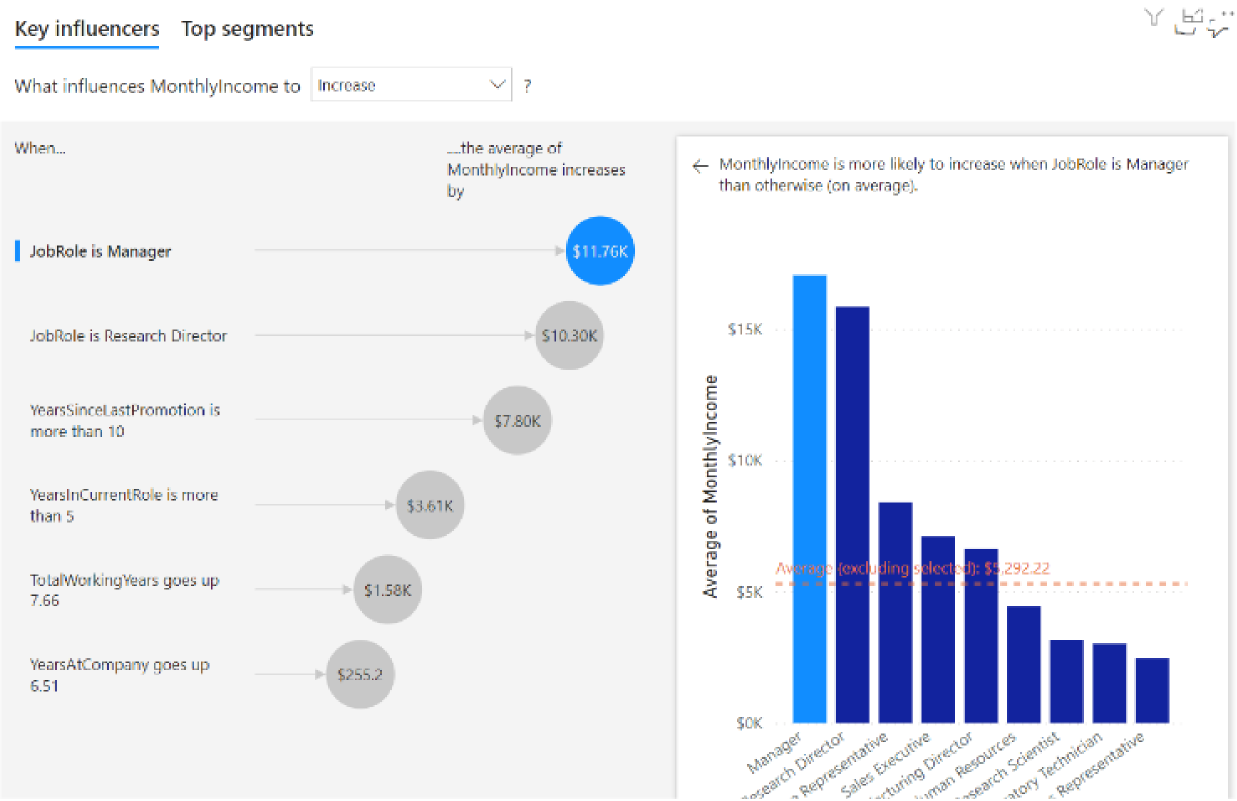
What are "key influencers"?



What are "key influencers"?



Key influencer visualization

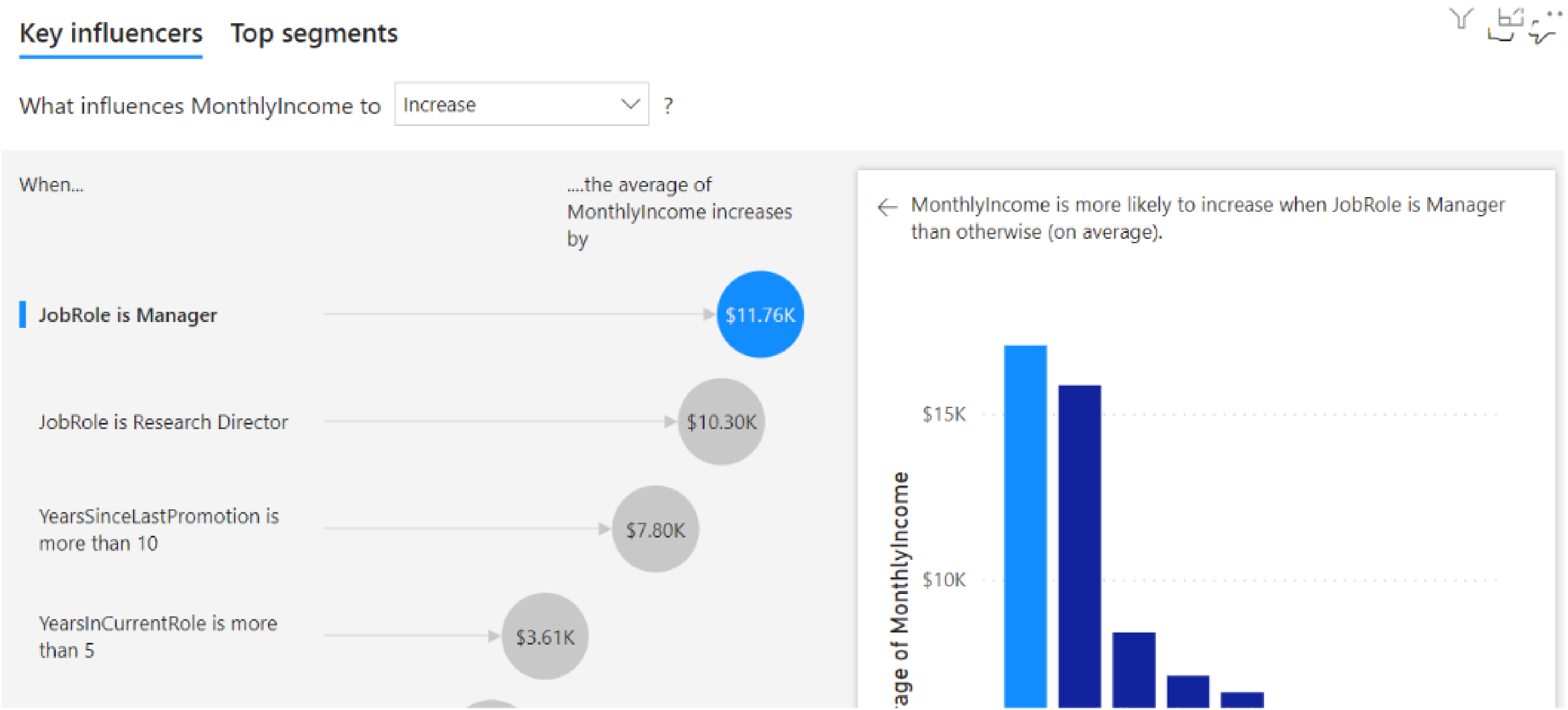


What influences MonthlyIncome to Increase ?

When...	...the average of MonthlyIncome increases by
JobRole is Manager	$11.76K
JobRole is Research Director	$10.30K
YearsSinceLastPromotion is more than 10	$7.80K
YearsInCurrentRole is more than 5	$3.61K
TotalWorkingYears goes up 7.66	$1.58K
YearsAtCompany goes up 6.51	$255.2

MonthlyIncome is more likely to increase when JobRole is Manager than otherwise (on average).

Key influencer visualization



What influences MonthlyIncome to Increase ?

When...

....the average of MonthlyIncome increases by

JobRole is Manager	$11.76K
JobRole is Research Director	$10.30K
YearsSinceLastPromotion is more than 10	$7.80K
YearsInCurrentRole is more than 5	$3.61K

MonthlyIncome is more likely to increase when JobRole is Manager than otherwise (on average).

Reasons why key influencers are not found

- Problem: explanatory fields have too many categories

Action: transform variable into categories

- Problem: there aren't enough observations to derive patterns

Action: increase the sample size or find new data

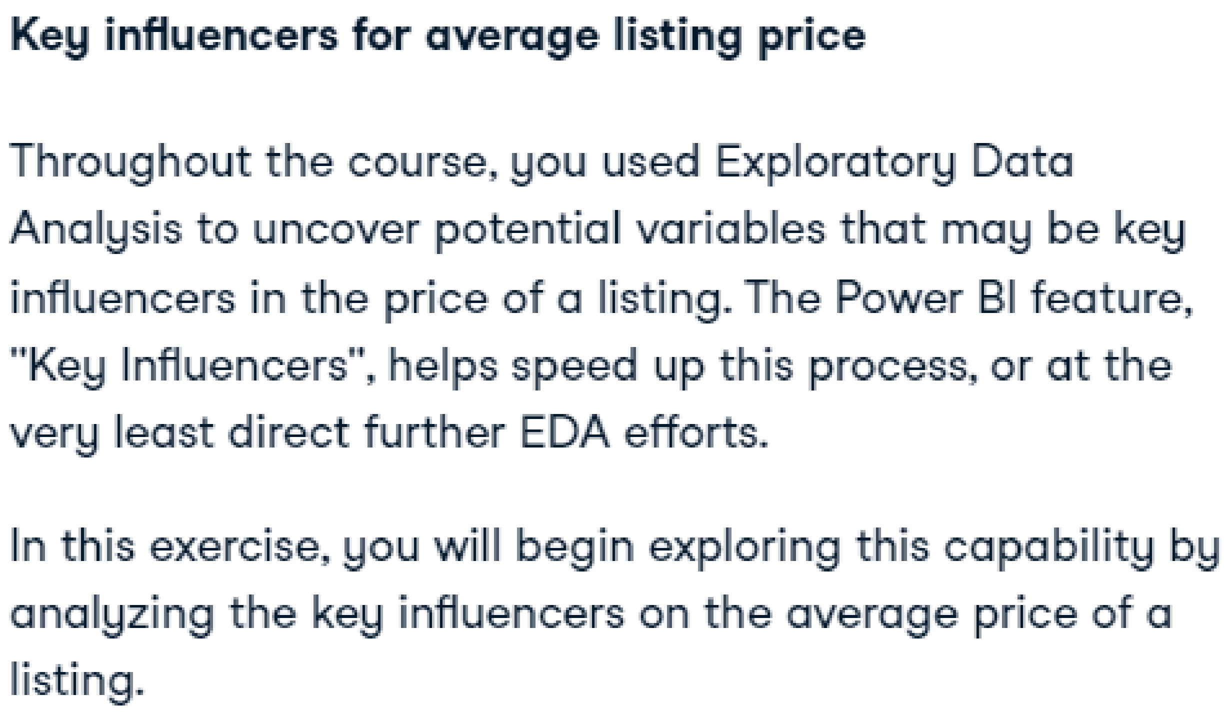
- There are simply no key influencers

Action: explore other explanatory variables

"Decomposition trees vs. key influencers"

The two visualizations introduced in this lesson - Decomposition Trees and Key Influencers - are both useful for analyzing a target variable by other explanatory variables. However, it's important to understand their difference and use cases.

"Complete Exercise"

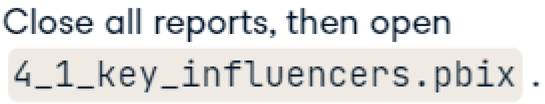


Key influencers for average listing price

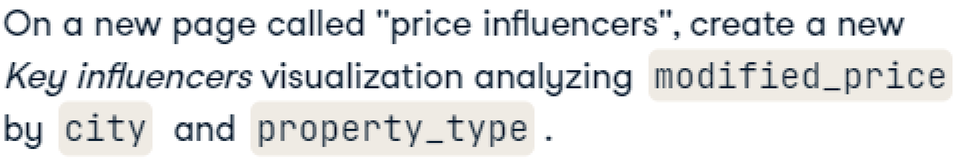
Throughout the course, you used Exploratory Data Analysis to uncover potential variables that may be key influencers in the price of a listing. The Power BI feature, "Key Influencers", helps speed up this process, or at the very least direct further EDA efforts.

In this exercise, you will begin exploring this capability by analyzing the key influencers on the average price of a listing.

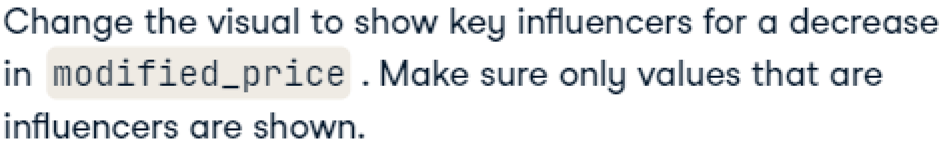
“Instructions"



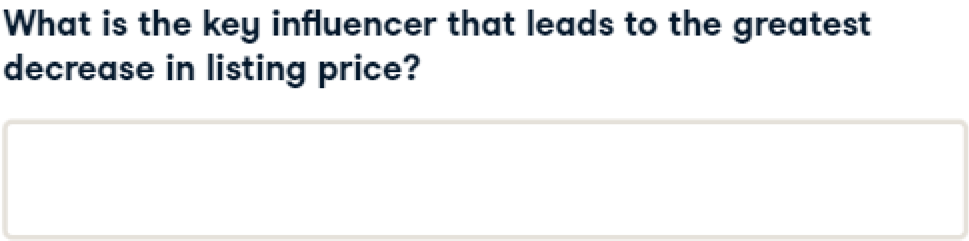
Close all reports, then open `4_1_key_influencers.pbix` .



On a new page called "price influencers", create a new *Key influencers* visualization analyzing `modified_price` by `city` and `property_type` .

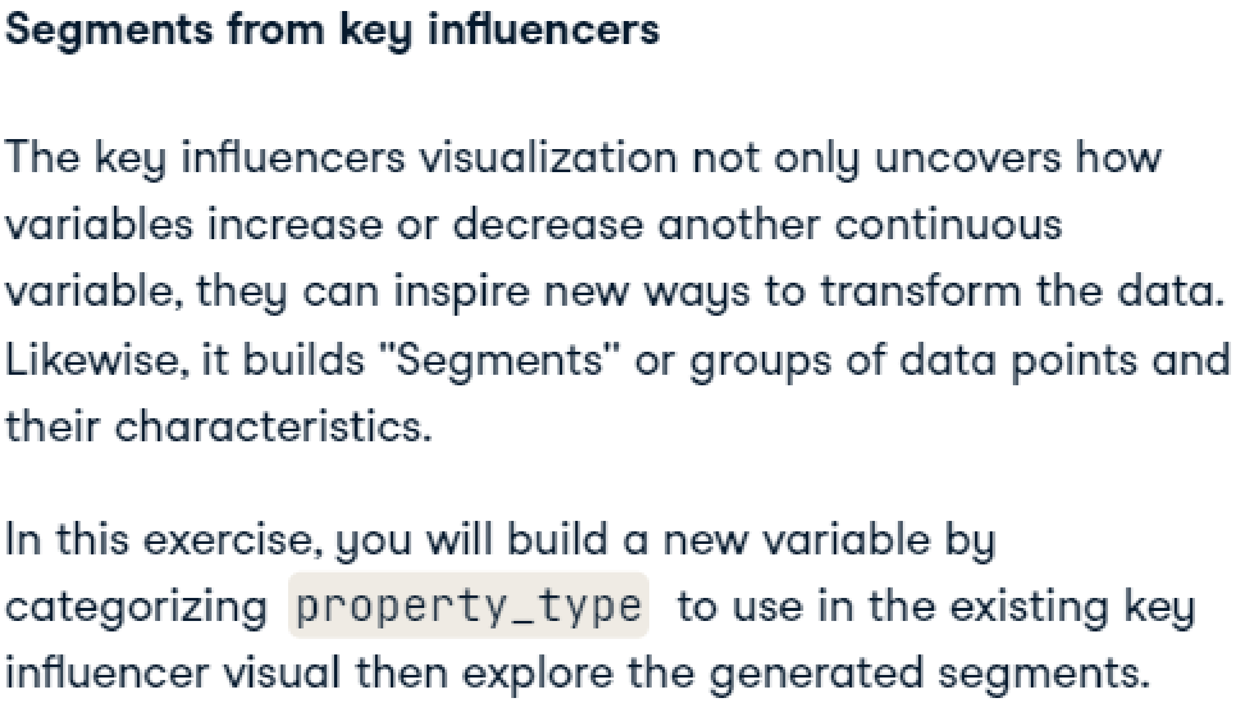


Change the visual to show key influencers for a decrease in `modified_price` . Make sure only values that are influencers are shown.



What is the key influencer that leads to the greatest decrease in listing price?

"Complete Exercise"

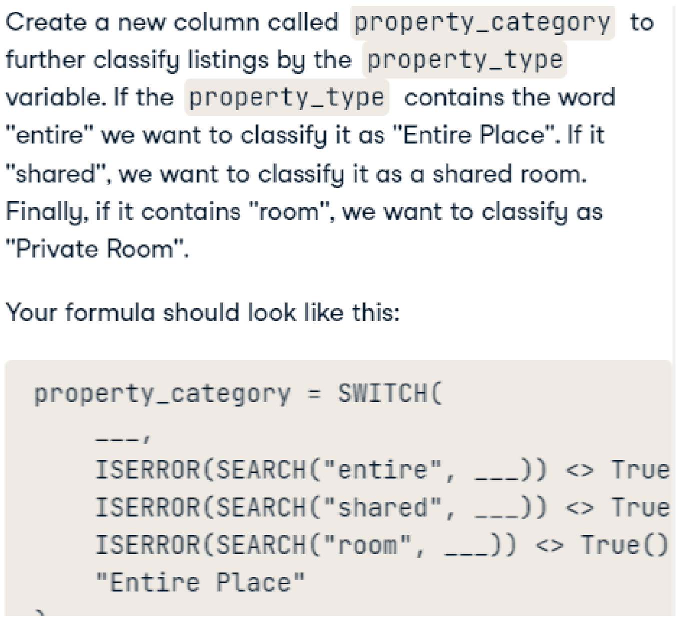


Segments from key influencers

The key influencers visualization not only uncovers how variables increase or decrease another continuous variable, they can inspire new ways to transform the data. Likewise, it builds "Segments" or groups of data points and their characteristics.

In this exercise, you will build a new variable by categorizing `property_type` to use in the existing key influencer visual then explore the generated segments.

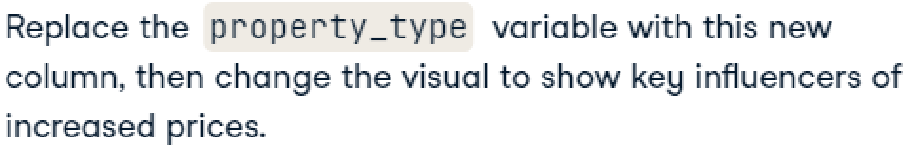
“Instructions"



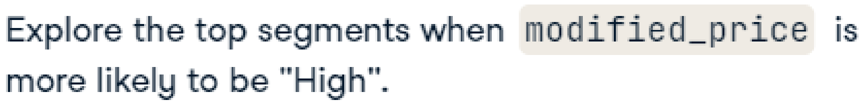
Create a new column called `property_category` to further classify listings by the `property_type` variable. If the `property_type` contains the word "entire" we want to classify it as "Entire Place". If it "shared", we want to classify it as a shared room. Finally, if it contains "room", we want to classify as "Private Room".

Your formula should look like this:

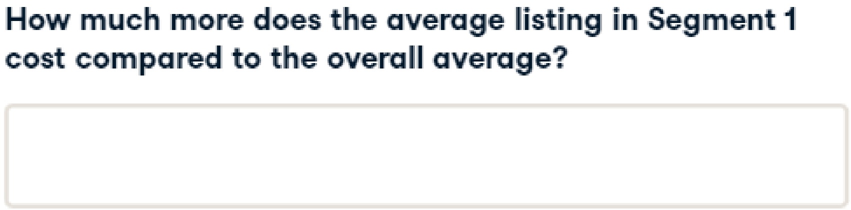
```
property_category = SWITCH(
    ___,
    ISERROR(SEARCH("entire", ___)) <> True
    ISERROR(SEARCH("shared", ___)) <> True
    ISERROR(SEARCH("room", ___)) <> True()
    "Entire Place"
)
```



Replace the `property_type` variable with this new column, then change the visual to show key influencers of increased prices.

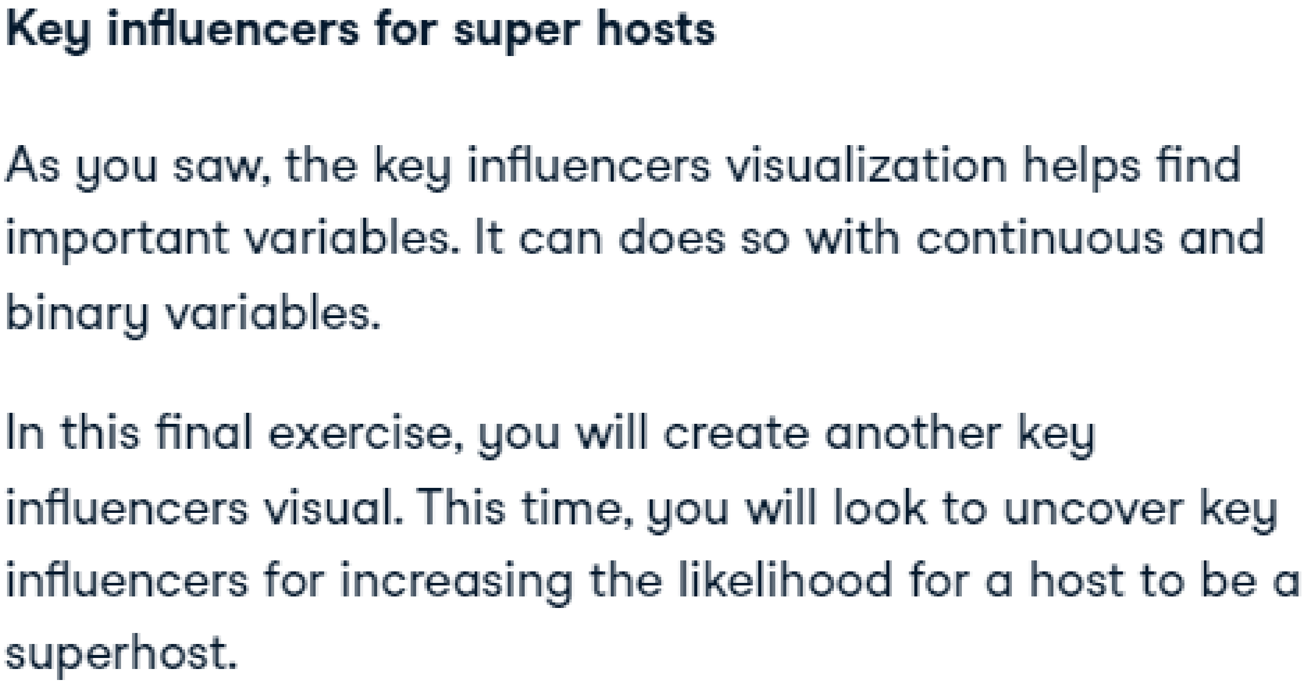


Explore the top segments when `modified_price` is more likely to be "High".



How much more does the average listing in Segment 1 cost compared to the overall average?

"Complete Exercise"

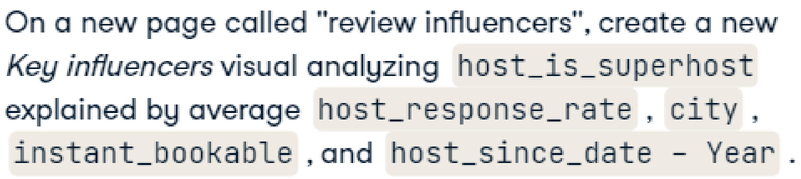


Key influencers for super hosts

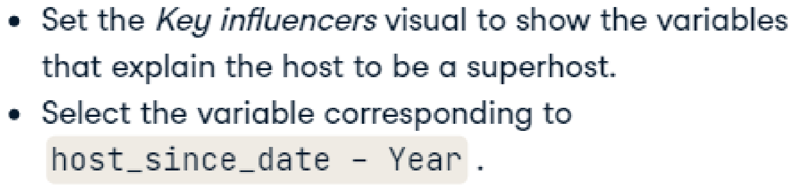
As you saw, the key influencers visualization helps find important variables. It can does so with continuous and binary variables.

In this final exercise, you will create another key influencers visual. This time, you will look to uncover key influencers for increasing the likelihood for a host to be a superhost.

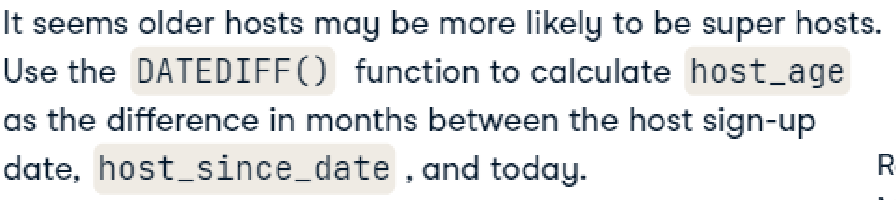
“Instructions"



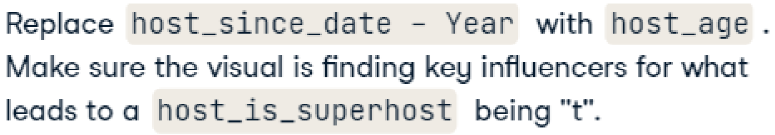
On a new page called "review influencers", create a new *Key influencers* visual analyzing `host_is_superhost` explained by average `host_response_rate`, `city`, `instant_bookable`, and `host_since_date - Year`.



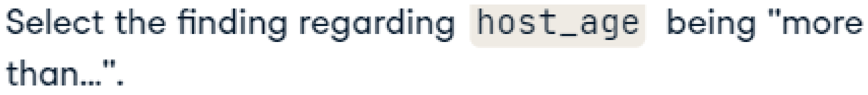
- Set the *Key influencers* visual to show the variables that explain the host to be a superhost.
- Select the variable corresponding to `host_since_date - Year`.



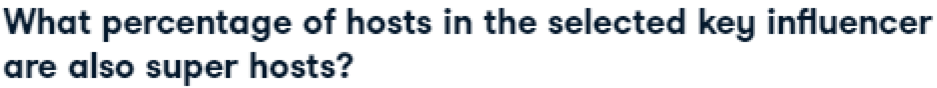
It seems older hosts may be more likely to be super hosts. Use the `DATEDIFF()` function to calculate `host_age` as the difference in months between the host sign-up date, `host_since_date`, and today.



Replace `host_since_date - Year` with `host_age`. Make sure the visual is finding key influencers for what leads to a `host_is_superhost` being "t".



Select the finding regarding `host_age` being "more than...".



What percentage of hosts in the selected key influencer are also super hosts?