# Radio Station Challenge

Learning Objectives

### To understand and experience the industrial best practices for exploratory data analysis

* To understand the common causes that could lead to poor quality of analysis --- and therefore how to avoid them
* To conduct exploratory data analysis using python

The Task

This activity involves starting a radio station focusing on the young people in Slovakia[[1]](#footnote-1). Your organisation’s mission is to be a leading radio entertainment provider in the country for young people below the age of 30, given this is by far the biggest segment of listeners tuning into your station.

Due to budget and time constraints, you were unable to perform a tailored customer survey, but luckily, on kaggle ([www.kaggle.com](http://www.kaggle.com/)) a survey of the personal preferences and living habits of over 1010 young people in Slovakia was carried out in 2013.

While your organisation currently has a limited budget (if any) to develop a sophisticated data warehouse, your Director of Information (who is double-hatted as Director of People and has a HR background) believes there are opportunities to develop a data analytics capability in the future.

Your team’s task is to understanding how to form the relevant questions you wish to ask that would enable you to achieve the business’s goal of increasing the number of listeners.

More specifically:

How to find answers to those questions from the data you have

How you can be sure the answers obtained are correct

How decisions can be made based on those answers that would influence the achievement of the strategic objective.   
  
Furthermore, your team should also recommend what improvements can be made to the analytic process in the future, if your company develops its own data analytics capability.

Background

The young person’s survey data can be found in the challenges folder or from the link below:

<https://www.kaggle.com/miroslavsabo/young-people-survey>

In 2013, students of the Statistics class at [FSEV UK](https://fses.uniba.sk/en/)  (Faculty of Social and Economic Sciences, The Comenius University in Bratislava) were asked to invite their friends to participate in this survey. 1010 young people, all of whom were of Slovakian nationality, aged between 15-30, took part.

150 questions were asked in the survey and can be split in the following groups:

* Music preferences (19 items)
* Movie preferences (12 items)
* Hobbies & interests (32 items)
* Phobias (10 items)
* Health habits (3 items)
* Personality traits, views on life, & opinions (57 items)
* Spending habits (7 items)
* Demographics (10 items)

The answers can be classified into three types:

* Ratings: from 1 to 5. So for example, on music preference: “Pop” would refer to the question: “Do you like Pop music?”, and rating of 1 means “very much dislike” and 5 means “very much like”; on Phobias, “Spider” would refer to the question “Do you fear spiders?”, and 1 would mean “not at all” and 5 would mean “fears a lot”.
* Categories: these are non-numerical values, so “Village - Town” would refer to the question: “Do you live in a village or a city?”, and answer would either be “village” or “city”. Similarly, for gender, males will be “male” and females “female” and so on.
* Numerical values: the exact values of age, height, weight etc.

The data, consisting of all the responses collected from the 1010 participants, is stored as an CSV (comma separated value) file that can be opened in Microsoft Excel.

The detailed explanation of each of the questions asked and what column of the CSV file corresponds to which question can be found on the kaggle page by following the link given above.

The data contain missing values i.e. the young people did not answer all the questions. It also contain ambiguous answers, most likely due to careless (inevitable) mistakes made when the organisers entered the data from paper survey forms collected.

As stated already, due to budget and time constraints, there will be no extra data available for your team’s initial analysis. So all the analysis and conclusions you draw will have to be based on this data that we have to hand.

Detailed requirements

In its simplest terms, your task can be broken down into 3 parts

# Activities

## Part 1 – Business Understanding

Understand how to convert the strategic goal of increasing the listeners to your radio station into a set of detailed and answerable questions that can be ‘presented’ to the data, and how the stakeholder should utilise the possible answers in deciding the radio programmes.

You may want to think about the following (N/B. This is not a complete or definitive list):

* + How would genre affect your station’s ratings
  + What happens if a person tuning in and hears a song he loves?
  + What happens if a person tuning in and hears a genre he hates?
  + What about talk shows?
  + What topics would one broadcast that may increase popularity?
  + Would the popular topics be related to particular genre preferences of the listeners?
  + Do you want to target a specific demographic? E.g. Should you target music lovers, and ignore people who are indifferent?
  + Use your intuition, can the data answer a question posed by your team?

For this part, you may want to read the CRISP-DM document: <https://www.the-modeling-agency.com/crisp-dm.pdf>, Chapter I and Chapter II-1: Business Understanding.

## Part 2 – Exploratory Analysis

Understand why the correct understanding of your data is crucial to any data analytic undertaking. You should find out:

* + How does the size of your data sample affect the final outcome of the analysis?
  + What affects do missing data have on your results?
  + How to deal with missing data?
  + How to deal with outliers?
  + What is the distribution of Discrete/Categorical or Continuous data?
  + What factors would affect your confidence on the results of the analysis?
  + What are the correlations within the data set?

In this part, you will need to examine data and perform exploratory data analysis using python.

If you have time, you can create Plotly charts to enhance data exploration and analysis.

Activities (cont.)

## Part 3 - Explanatory Analysis

The results from your analysis should be evaluated in the context of your strategic goal. In addition, you may want to comment on

* + What further questions you may want to ask in light of the insights obtained so far.
  + What are the assumptions being made in the analysis?
  + Are the assumptions made consistent with the insights we have obtained?

You can also use [data story telling](https://venngage.com/blog/data-storytelling/) techniques to give a presentation on your recommendations to the business as to how to realise the strategic goal and the reasons of why you have reached those recommendations backed up by data evidence.

Think perhaps of a triangle. At the top is your decision, in the next layer your reasons why and the layer underneath the source data.

Try to avoid presenting the source data first. Rather present the decisions and then how you have reached those and what source/raw data they are based on. You might like to think about presenting the chances that you could be wrongly interpreting the data or how sure you can be that the data is representing what you need it to represent.

Useful links:

Matplot Lib:

[matplotlib/AnatomyOfMatplotlib: Anatomy of Matplotlib -- tutorial developed for the SciPy conference (github.com)](https://github.com/matplotlib/AnatomyOfMatplotlib)

Seaborn Gallery

<https://seaborn.pydata.org/examples/index.html>

Plotly

<https://plot.ly/python/widget-app/>

IPywidgets

<https://towardsdatascience.com/interactive-controls-for-jupyter-notebooks-f5c94829aee6>

<https://towardsdatascience.com/interactive-data-analysis-with-dropdown-menu-ipywidgets-and-plotly-in-jupyter-notebook-591a84a81b22>

Data Storytelling

<https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/>

https://venngage.com/blog/data-storytelling/

Visualisation tips

<https://www.elsevier.com/connect/a-5-step-guide-to-data-visualization>

1. If you are wondering why Slovakia? That is simply because this is the nationality of all the participants of the survey data we are using for this workshop. Which country we choose is irrelevant to the purpose of this workshop. [↑](#footnote-ref-1)