Radio Station: can we increase the number of listeners?

DATA

In 2013, students of the Statistics class at FSEV UK (Faculty of Social and Economic Sciences, The Comenius University in Bratislava) were asked to invite their friends to participate in a survey. 1010 young people took part. All of the participants were of Slovakian nationality and aged between 15-30.

The survey was made up of 150 questions, which can be split into 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:

1. **Ratings**: From 1 to 5. So for example, in the 'music preferences' category, 'Pop' would refer to the question, "Do you like Pop music?" and a rating of 1 means 'very much dislike' and 5 means 'very much like'.
2. **Categories**: These are non-numerical values, so 'Village – Town' would refer to the question, "Do you live in a village or a city?", and the answer would either be 'village or 'city'.
3. **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 a CSV (Comma Separated Value) file.

You can find a detailed explanation on the Kaggle page of each of the questions asked and what column of the CSV file corresponds to which question.

The data contains 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.

There is no extra data available for your team's initial analysis. You will have to base all analyses on the data you have to hand.

Part One

You need to 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. You also need to think about how the stakeholder should utilise the possible answers in deciding the radio programmes.

You may want to think about the following:

- How would genre affect your station's ratings?
- What happens if a person tunes in and hears a song they love?
- What happens if a person tunes in and hears a genre they hate?
- What about talk shows?
- What topics could you 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?
- Can the data answer a question posed by your team?
    - EXTRA reading: Chapter I & II-1 of **CRISP-DM guide**

Part Two

You need to understand why the correct understanding of your data is crucial to any data analytic undertaking.

Try to answer the following:

- 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 factors would affect your confidence on the results of the analysis?

In this part, you'll need to examine data and perform exploratory data analysis using Python.

EXTRA: use interactive IPython widgets and Plotly

Part Three

You should evaluate the results from your analytics team in the context of your strategic goal. In addition, you may want to comment on the following:

- What further questions may you 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 that you have obtained?

You should use **data story telling techniques** to give a presentation to the business, focused on your recommendations as to how to realise the strategic goal and the reasons behind why you have reached those recommendations. You should back this up with data evidence.

SUGGESTED structure of presentation:

1. Your decision/recommendation.
2. How you have reached your decision.
3. The source/raw data evidence that supports your decision.

You might also like to talk about the possibility that you could be wrongly interpreting the data, or how sure you are that the data is representing what you need it to represent.