Here are the revision notes based on the class session about data analysis:

Data Analysis Process and Techniques

In this session, we explored the fundamental steps of performing a data analysis task and the various techniques applicable when handling different types of data. Here's a detailed breakdown:

1. Problem Statement Definition

The first step is to clearly define the problem statement. This involves explaining the problem in your own words, which gives you a deeper understanding and acts as a guide throughout the analysis process. In the context of the Netflix case study discussed, the problem statement was to analyze data to help Netflix decide which type of shows or movies to produce and how to grow their business in different countries【6:1†source】.

2. Data Loading and Initial Exploration

- **Loading the Data**: Import necessary libraries (such as Pandas, NumPy) and load your dataset using functions like `pd.read_csv()` for CSV data【6:13†source】.
- **Initial Checks**: Check the shape of the dataset to know the number of rows and columns using `df.shape`. Verify column types, and get a summary of non-null values using `df.info()` 【6:0†source】【6:6†source】.
- **Null Values**: Determine the number of null values in each column with `df.isnull().sum()`, and calculate the percentage of null values if necessary【6:0†source】.

3. Data Analysis

Univariate and Bivariate Analysis

- **Bivariate Analysis**: Employ scatter plots and correlation matrices to analyze relationships between two features. For example, analyzing the correlation between the release year and duration to observe trends 【6:3†source】.

Categorical and Numerical Features

- **Categorical Features**: Understand the nature of categorical attributes, which are discrete and signify categories like genre, type, rating, etc. These need conversion using methods like `.astype('category')` 【6:14†source】.
- **Numerical Features**: Analyze continuous features (e.g., duration, age) that typically require statistical summaries using `df.describe()` 【6:7†source】【6:14†source】.

Data Transformation

- **Exploding Columns**: For columns with list-like entries, such as genres that contain multiple categories per row, use `.apply(lambda x: x.split(','))` followed by `df.explode(column_name)` to convert these into separate rows for each list entry【6:10†source】【6:19†source】.

4. Handling Missing Data and Outliers

- **Missing Data**: Although optional at this stage, outlier and missing data treatment is crucial, using strategies such as imputation or deletion based on IQR (Interquartile Range)【6:16†source】【6:12†source】.
- **Outliers**: Identify outliers by analyzing data with box plots. Apply IQR-based formulas for defining acceptable ranges and filter data accordingly 【6:5†source】【6:16†source】.

5. Insights and Conclusion

The session emphasized the extraction of insights from the analysis conducted. Observations on customer preferences based on data exploration, such as the popularity of specific movie genres in different regions, are critical for making informed business decisions【6:11†source】【6:18†source】. These insights can lead to

Submitting Analysis: Utilize platforms like Google Colab for computations and markdown for documentation. Convert and submit the analysis as a PDF document【6:4†source】【6:18†source】.

This guide provides a structured approach to performing data analysis and is aimed at ensuring robust practices for extracting meaningful insights from datasets.