**Space Mission Data Analysis Project Report**

1. Introduction

This project was created as part of a data analysis course, with the primary goal of applying and evaluating the techniques we learned during the course's duration. Rather than working on familiar datasets like finance, hospitals, or sales, we purposefully chose a distinct and uncommon topic: space missions.

The goal was to design and create a complete data process, which included database schema design and data generation, data cleaning and logical validation, and finally complex analysis and interactive dashboards. By selecting a challenging dataset with various entities and relationships, we were able to practice database management, SQL analysis, Python-based preprocessing, and dashboard creation with tools like Excel, Power BI, and Streamlit.

1. Database Design

We began by creating a relational database schema that represented different aspects of space exploration. The database included five main tables:

* **Agencies**: agency information (ID, name, country, founded year, budget).
* **Missions**: mission details (ID, name, agency, destination, type, status, launch date, duration).
* **Spacecraft**: spacecraft attributes (ID, name, type, cost, manufacturer, first launch year, linked mission).
* **Astronauts**: astronaut demographics and career statistics (ID, name, nationality, gender, birth year, mission count).
* **Mission\_Crew**: mapping of astronauts to missions, including their roles.

An Entity-Relationship Diagram (ERD) was created to visually represent the schema and guarantee that all foreign key relationships were appropriately defined.

In order to fill the database, Python's Faker library was utilized to create fake data. However, because Faker generates random data without logical consistency, we expected to need significant changes. For example, without disruption, we could easily achieve unbelievable records such as a 25-year-old astronaut completing six missions or Moon missions lasting longer than Mars missions.

1. Data Generation and Cleaning

After the initial CSV files were prepared, the data was modified several times to verify logical consistency.

**Database & SQL**

* Built the SQL Server database from the CSV files.
* Documented the schema and performed query-based analysis directly in SQL Server.
* Added more realistic data to the database and added a new status to the missions tables (Planned status).

**Python Cleaning & Adjustments**

* Increased the number of astronauts to match the number of missions and ships.
* Astronauts were assigned realistic mission participation, ensuring that the number of missions was appropriate for their age and level of expertise.
* The mission crew table was enforced to include a particular number of astronauts per mission, ensuring that each mission had an acceptable number of participants with consistent roles.
* Adjusted mission statuses: the original dataset included Planned, Ongoing, Completed, and Failed. The Ongoing status created significant logical complications with dates (for example, future missions with already allocated launch dates). To address this, ongoing missions were replaced with new Completed and Failed statuses to ensure time consistency.
* Corrected mission durations: for example, Mars journeys are guaranteed to be longer on average than Moon missions.

These changes changed the artificial dataset from a completely random structure to one that reflected more realistic relationships and restrictions found in real-world space exploration data.

1. Data Analysis

**SQL Analysis:** SQL queries were utilized to obtain descriptive, relational, and KPI-style information.

* Descriptive queries include: missions by agency, astronauts by nationality, gender distribution, average mission duration by destination, and spacecraft type distribution.
* Relational queries: linking agencies to missions, astronauts to missions, and spacecraft to mission allocation.
* KPIs: longest mission, oldest astronaut in space, most frequently used spacecraft, astronaut with the highest mission count, and the agency with the largest budget.

**Python Preprocessing:** The main goal was to alter the raw data to remove logical inconsistencies before producing the CSV files for further analysis in dashboarding tools.

1. Dashboards

**Excel Dashboard:** An interactive dashboard was developed in Excel, focusing on astronaut data:

* Filters for gender, nationality, and birth year.
* Gender distribution pie chart.
* Bar chart of astronauts per country.
* World map showing distribution of missions by country.
* Top 10 astronauts ranked by mission participation.

**Power BI Dashboards:** Four Power BI dashboards were created to analyze:

1. **Missions**: Mission growth over time, mission statuses per agency, and comparisons across destinations.
2. **Agencies**: Mission counts per agency, geographic distribution, and outcome success rates.
3. **Spacecraft**: types, cost trend, manufacturers' market share.
4. **Astronauts**: Demographics (age, gender, nationality), experience levels, and mission participation, Gender distribution.

**Streamlit Dashboard:** We developed an interactive Streamlit dashboard using Plotly Express to visualize and explore the dataset. The application featured a sidebar with filters for agency, mission type, and status, allowing all charts to update dynamically. The dashboard was organized into multiple tabs covering mission, agency, spacecraft, mission crew, and time analyses, enabling users to explore patterns such as first missions per agency, mission durations, and overall trends in space exploration. This dashboard provided an integrated and visually appealing approach to analyzing data from many angles.

1. Key Insights

From the combined analysis, several interesting results were obtained:

* **Mission Growth**: Significant rise after 2000, reflecting global expansion in space activity.
* **Destinations**: Majority to Low Earth Orbit and Sun-Synchronous Orbit; fewer to Moon and Mars.
* **Mission Types**: Satellite deployment and Earth observation dominate; fewer manned and deep-space missions.
* **Agencies**: NASA, Roscosmos, and ESA lead in missions; smaller agencies show more varied success rates.
* **Astronauts**: 600 total, average age 57, about 3 missions each. Genders are almost equally distributed.
* **Countries**: USA, Russia, and China have the most astronauts; others include Saudi Arabia, Chile, and Japan.

1. Challenges and Solutions

The artificial nature of the dataset was the most significant barrier to this project. Because Faker provides random but logically inconsistent data, we encountered numerous absurd scenarios. These featured astronauts with unrealistic career paths, mission statuses that clashed with launch dates, and destination-related durations that made no sense.

We were able to solve these issues by carefully applying logical rules in Python, replacing incorrect statuses, and matching values across tables, resulting in a consistent base for analysis.

1. Conclusion and Future Work

This project allowed us to exercise and combine every aspect of data analysis skills learned during the course. We designed and implemented a relational schema, generated and cleaned fake data, performed SQL-based analysis, and created dashboards in Excel, Power BI, and Streamlit.  
Beyond technical skills, the project highlighted the value of data quality and logical structure in analysis. Even with advanced tools, incorrect data leads to false insights, which we had to actively address.

If we continue this project, further work may include:

* Using real-world datasets (such as NASA open data).
* Dashboards are being improved to include more complex features.
* Enhancing functionality in Streamlit for storytelling and simulation scenarios.

Overall, the project met its goal of applying course principles to a unique and complex topic while also showcasing our ability to handle the entire data analysis process, from raw input to interactive insight delivery.