**Project Summary**

**Data Dictionary**

The project involves the creation of a database following the Star Schema format, consisting of the fact table Student and four dimension tables: Decoded, Survey, TestScore, and Demographics. The Data Dictionary outlines the structure of these tables, including primary and foreign keys, data types, and constraints.

**Data Cleaning and Preparation**

**Modifications to Raw Data:**

1. **Database Design:** Reorganized raw data into a Star Schema format, adding primary keys (TestScore\_ID, Demographics\_ID, Survey\_ID, Decoded\_ID) and modifying table structures accordingly.
2. **Fact Table Creation:** Introduced a Student table as the central Fact Table, integrating foreign keys from dimension tables. Removed redundancy by eliminating StudentIDs from dimension tables.
3. **Data Integrity:** Checked for duplicates and null values, ensuring original data integrity. Utilized Conditional Formatting and Sorting to identify and address any issues.
4. **File Formats:** Maintained consistency with file formats - CSV for MySQL and BigQuery imports, XLS for presentation purposes.
5. **Database Modification:** Faced a blocker related to reversed relationships. Modified the database by creating new tables, adding columns, and deleting redundant StudentID columns to adhere to Star Schema requirements.

**Exploratory Data Analysis**

**Visual Exploration:**

1. **Charts:** Explored raw data through charts, including gender distribution and parent education levels.
2. **Outlier Identification:** Used Box & Whisker plots to identify outliers in test scores, aiding in understanding potential influential points.
3. **Distribution Analysis:** Checked the distribution of test scores, relative frequency of students having lunch, completing test preparation, and cross-tabulations based on lunch type and test preparation.
4. **Descriptive Statistics:** Calculated various metrics using Excel's Descriptive Statistics function, highlighting mean, standard deviation, standard error, range, and count.
5. **Regression Analysis:** Conducted regression analysis on lunch plan, test preparation, and parent education against average test scores.

**MySQL Schema**

1. **Database Creation:** Established the MySQL schema, creating tables with appropriate constraints, defaults, and relationships.
2. **ERD Star Schema:** Generated an Entity-Relationship Diagram (ERD) depicting the Star Schema structure, making necessary adjustments.
3. **Data Import:** Imported data into MySQL, addressing errors and ensuring data integrity.

**BigQuery:**

1. **Jennifer's Query:** Average scores for each subject grouped by year group.
2. **Natalie's Query:** Combined average scores for each year group.
3. **Nora's Query:** Average scores for each subject.
4. **Louisa's Query:** Average scores across subjects grouped by gender.
5. **Aurora's Query:** Number of males and females in each year group and relative frequency.
6. **Aurora's Query (contd):** Average test scores grouped by parental level of education.
7. **Aurora's Query (contd):** Standard deviations for each subject.
8. **Nina's Query:** Did students who took the test prep course score higher? Does test preparation affect the scores?
9. **Nina's Query (contd):** Did the meal plan affect test scores?
10. **Nina's Query (contd):** Test scores grouped by both preparation and lunch.
11. **Victoria's Query:** Correlation coefficients of student test results based on students who had BOTH a standard meal plan AND completed test prep.
12. **Susana's Query:** Correlation of combined averages of test scores with BOTH completion of test prep AND standard meal.
13. **Maya's Query:** Combined average test scores grouped by parent level of education and conditions of having a standard lunch and completed preparation vs. none and free lunch.
14. **Ivy's Query:** Is there a correlation between parent education and meal plan?
15. **Hazel's Query:** Correlation between test preparation and parent education.
16. **Elena's Query:** Rate of meal purchase for parents of each education level joined with the rate of completing the test preparation.
17. **Elena's Query (contd):** Rate of meal purchase for parents of each education level.
18. **Elena's Query (contd):** Correlation coefficient for parent education and average math scores.
19. **Nina's Query (contd):** Is there a correlation between test prep (yes) and meal plan (standard)?

These queries cover a wide range of analyses, including average scores, correlations, rates, and groupings based on various conditions such as meal plans and test preparation. The queries provide insights into the relationships between different variables in the dataset, helping to understand the factors influencing student performance.

**Work Summary**

1. **Excel Work:** Created the fact table, added columns for each dimension table, and performed exploratory analysis.
2. **MySQL Work:** Established relationships, created tables, and imported data, ensuring constraints and data accuracy. Created a new table with decoded values and imported it into MySQL and BigQuery.
3. **GCP BigQuery:** Imported data into BigQuery, conducted joins, and analyzed data.
4. **Challenges:** Overcame challenges, such as modifying relationships and resolving import errors.
5. **PowerPoint:** Prepared PowerPoint presentation and presented it to an audience.