📋 Post-Session Notes: Revision - Real Time Scenarios, Hypothesis Building, Right Data Collection, and Data Cleaning

---

🧩 Key Concepts Covered

1. **Introduction to Smart Health Tracker Project**
2. **Data Collection and Feature Identification**
3. **Loading the Data and Initial Inspection**
4. **Identifying and Handling Missing Values**
5. **Data Visualization to Understand Feature Distributions**
6. **Formulating Hypotheses for Statistical Testing**
7. **Hypothesis Testing Using Ordinary Least Squares (OLS) Regression**
8. **Linear Regression Model and Evaluation**
9. **Summary and Implications on Model Selection**

🔎 Detailed Breakdown

1. Introduction to Smart Health Tracker Project

- Mini-project centered on a smart health tracker dataset (~30,000 entries, 11 health-related features).

- Features included: age, gender, heart rate, sleep hours, calorie intake, stress level, mood, steps, and activity types.

- Interactive exercise via Mentimeter to brainstorm what data should be collected; results showed common health metrics like heart rate, blood pressure, sleep cycles, step counts.

- Emphasized real-world challenge: Unlike standard datasets, real data collection requires careful planning and understanding of what features matter for the problem at hand.

2. Data Collection and Feature Identification

- Importance of selecting the right features: domain knowledge and user needs guide what data to collect.

- Real-world data rarely comes clean or fully ready — many decisions must be made upfront.

3. Loading the Data and Initial Inspection

- Demonstrated loading CSV data into a pandas DataFrame.

- Checked for missing values using `isnull()` and computed the percentage of missing data per feature.

- Found about 1% missing data overall, which was considered small enough to address without removing rows.

4. Identifying and Handling Missing Values

- **Discussed multiple strategies to handle missing data:**
    - **Dropping rows: simple but risks losing valuable data.**
    - **Imputation:**
        - **For numerical features (age, heart rate): mean or median imputation.**
        - **For categorical features (gender, mood): mode imputation.**
    - **Interpolation or predictive filling mentioned as advanced options.**
- **Chose to fill missing values to preserve dataset size and comply with ethical data use.**
- **For age, median preferred over mean due to outliers.**
- **For categorical variables, filled with the most frequent category.**

5. Data Visualization to Understand Feature Distributions

- **Various plots used to explore data distributions and spot anomalies:**
    - **Histograms: age, daily steps, sleep hours, calorie intake, stress level, mood.**
    - **Box plots: resting and active heart rate — helped identify outliers.**
    - **Bar charts: gender distribution, activity types.**
- **Key findings:**

- Age distribution skewed younger (mostly 20-40 years).
- Gender roughly balanced, minimizing sampling bias.
- Steps and sleep roughly normal distributions.
- Stress varied widely, centered near neutral.

- Visualization helped validate data quality and provided intuitive insights for further analysis.

6. Formulating Hypotheses for Statistical Testing

- Reviewed how to write null (H0) and alternative (H1) hypotheses.
- Example hypothesis:
  - H0: No significant relationship between daily steps, stress level, and hours of sleep.
  - H1: There is a significant relationship.
- Two-tailed test used since no direction of effect was assumed.

7. Hypothesis Testing Using Ordinary Least Squares (OLS) Regression

- Used `statsmodels` to perform OLS regression on the dataset.
- Key outputs focused on:
  - R-squared: Proportion of variance explained by model (~0 here → no explanatory power).

    - P-values: Statistical significance of predictors (values > 0.05 → not significant).

- Results accepted the null hypothesis: no evidence that daily steps or stress level significantly predicted hours of sleep.

8. Linear Regression Model and Evaluation

- Confirmed results with scikit-learn linear regression.
- R^2 score also near zero, supporting OLS findings.
- Regression line visualization showed no meaningful relationship.
- Suggested linear modeling insufficient, encouraging exploring nonlinear or more complex models later.

9. Summary and Implications on Model Selection

- Real-world data science pipeline involves:
    - Thoughtful data collection and feature selection.
    - Rigorous data cleaning and missing data handling.
    - Exploratory data visualization to understand data properties.
    - Formulating and testing statistical hypotheses before modeling.
- Found that linear relationships might not exist in this dataset, signaling the need for alternative models or additional variables.

- This incremental approach ensures models built are based on sound understanding rather than assumptions.

- The project will continue building on these foundations with more advanced techniques.

📌 Key Takeaways

- Data collection is critical: Knowing what and how to collect data affects all downstream tasks.

- Handling missing data carefully avoids loss of valuable information and respects data ethics.

- Visual exploration is an essential sanity check and provides initial insights.

- Hypothesis testing informs whether variables are related, helping avoid wasted effort on irrelevant features.

- Model evaluation metrics (R^2, p-values) help judge model validity and fit.

- Iterative and evidence-based approaches outperform blind model building.

🔗 Real-World Application Context

- The smart health tracker use case represents many IoT and wearable device data science challenges.

- **Handling imperfect data is a norm, not an exception.**
- **Understanding domain and user needs helps guide effective feature engineering.**
- **Statistical rigor provides confidence in model-driven decisions impacting health and wellbeing.**
