**Project Title:** Mental Health Impact of Remote Work

**Team Members:** Daria Savtšenko, Anton Voronkov, Elisabeth Serikova

**Project repository:** https://github.com/daria-sav/DS-RemoteWork-Wellbeing.git

## Task 2. Business understanding

### 1. Identifying Business Goals

**Background:**

Remote work has become a norm across industries, particularly due to the COVID-19 pandemic, leading to a significant shift in workplace dynamics. While it has provided flexibility and reduced commuting, remote work has also posed challenges to employees' mental health, such as increased stress, isolation, and difficulty balancing work and personal life. Understanding these impacts is crucial to developing actionable strategies for improving remote work conditions and promoting mental well-being among employees.

**Business Goals:**

- Identify the key factors influencing the mental health of remote workers, such as workload, social support, and work-life balance.
- Develop actionable recommendations for employers to mitigate stress and prevent mental health issues in remote employees.
- Understand the demographics and professional groups that are most and least satisfied with remote work, enabling tailored interventions.

**Business Success Criteria:**

- Deliver a comprehensive report that outlines findings and provides actionable recommendations for employers and HR teams.
- Create clear visualizations that illustrate the correlation between stress levels and remote work conditions across different professions and demographics.

### 2. Assessing the Situation

**Inventory of Resources:**

- **Data:** Kaggle dataset on remote work and mental health (link), consisting of 14 MB of data including demographic information, remote work conditions, and mental health metrics.
- **Technologies:** Python programming tools (Pandas for data analysis, Matplotlib/Seaborn for visualization, and Scikit-learn for machine learning tasks), Jupyter Notebook for workflow management.

**Requirements, Assumptions, and Constraints:**

- **Requirements:**

- Ensure thorough cleaning and preprocessing of the dataset to handle missing or inconsistent data.
- Adhere to ethical data practices, ensuring no personal information is exposed in analysis or visualizations.
- Completion of the project by December 2024.

- **Assumptions:**
- Data is representative of a wide range of professions and demographics.
- Survey responses reflect accurate self-reported experiences.

- **Constraints:**
- Limited to the Kaggle dataset unless additional datasets can be incorporated within the timeline.
- Potential biases due to limited or uneven distribution of survey participants.
- Possible limitations arise from the self-reported nature of the data, which may introduce subjectivity and socially desirable responses. Furthermore, certain regions or industries might be underrepresented in the dataset.

**Risks and Contingencies:**

- **Risk 1:** Insufficient data quality or completeness.

**Contingency:** Supplement with additional datasets or restrict the analysis to high-quality subsets.

- **Risk 2:** Overlapping variables or unclear correlations.

**Contingency:** Use advanced feature engineering and statistical methods to refine the analysis.

**Terminology:**

- **Remote work:** Employment performed outside of a traditional office setting, typically from home.
- **Mental health:** Psychological well-being, encompassing stress, anxiety, and overall satisfaction.
- **Stress levels:** Self-reported indicators of perceived stress due to workload, isolation, or other factors.

**Costs and Benefits:**

- **Costs:**
- Team effort: Approximately 30–40 hours per member for data analysis, visualization, and reporting.
- Computing resources: Personal computers and open-source software tools (no additional costs).

- **Benefits:**

- Practical recommendations for employers to improve workplace conditions and employee satisfaction.
- Insights into the challenges and advantages of remote work, applicable to various industries.

**3. Defining Data-Mining Goals**

**Data-Mining Goals:**

1. Analyze and quantify the key factors influencing remote workers' mental health, such as work hours, communication frequency, and personal circumstances.
2. Cluster employee groups based on satisfaction levels to identify those needing additional support.
3. Create visualizations that demonstrate trends and correlations in stress levels across various professions and demographic groups.

**Data-Mining Success Criteria:**

- **Quantitative Metrics:**
  - High model accuracy (e.g., >85% for predictive models) in identifying stress-related factors.
  - Clustering analysis that provides clear and interpretable results, distinguishing groups effectively.
- **Qualitative Metrics:**
  - Visualizations that clearly illustrate insights and are easy to interpret for non-technical stakeholders.
  - Actionable findings that align with the stated business goals and support meaningful interventions.

**Task 3. Data understanding**

Task 3: Data Understanding

**Gathering Data**:

**Outline Data Requirements**

The objective of this project is to analyze the impact of remote work on mental health. For this, the data must include information related to:

Demographics: Age, gender, region.

Professional details: Job role, industry, years of experience, work location.

Work metrics: Hours worked per week, number of virtual meetings.

Mental health and well-being: Stress level, mental health condition, social isolation rating, work-life balance rating, access to mental health resources.

Additional factors: Sleep quality, physical activity, productivity change, satisfaction with remote work, company support for remote work.

**Verify Data Availability**

The dataset available from Kaggle contains 14 MB of relevant data, fulfilling the requirements. This dataset includes the columns specified in the project's background, ensuring availability of the necessary variables for analysis.

**Define Selection Criteria**

Include data points with valid and non-missing values for critical columns: "Stress_Level," "Mental_Health_Condition," and "Work_Life_Balance_Rating."

Focus on respondents engaged in remote work (filter rows where "Work_Location" equals "Remote" or similar).

Ensure a balance in representation across demographic and professional groups for meaningful insights.

Exclude data points with unrealistic or extreme outliers (like: 100+ hours worked per week).

**Describing Data**

The dataset consists of the following columns:

Employee_ID: Unique identifier for each respondent.

Age: Age of the respondent.

Gender: Self-identified gender.

Job_Role: The professional role of the respondent.

Industry: The industry in which the respondent is employed.

Years_of_Experience: Total professional experience in years.

Work_Location: Indicates whether work is remote or hybrid.

Hours_Worked_Per_Week: Average weekly working hours.

Number_of_Virtual_Meetings: Average number of virtual meetings attended weekly.

Work_Life_Balance_Rating: Self-assessed rating of work-life balance (scale 1-5).

Stress_Level: Self-reported stress level (scale 1-5).

Mental_Health_Condition: Self-reported presence of a mental health condition (Yes/No).

Access_to_Mental_Health_Resources: Indicates access to workplace mental health support (Yes/No).

Productivity_Change: Change in productivity (percentage increase/decrease).

Social_Isolation_Rating: Rating of social isolation (scale 1-5).

Satisfaction_with_Remote_Work: Overall satisfaction with remote work (scale 1-5).

Company_Support_for_Remote_Work: Self-assessed rating of company's support (scale 1-5).

Physical_Activity: Frequency of physical activity (like: daily, weekly).

Sleep_Quality: Self-reported sleep quality (scale 1-5).

Region: Geographic region of the respondent.

**Exploring Data**:

**Key Findings**

Demographics: Initial exploration shows balanced representation across genders but potential skew in some regions or industries.

Work Metrics: Most respondents report working 30-50 hours per week. Outliers include individuals working more than 70 hours weekly.

Mental Health: Stress levels and social isolation ratings show variability, with a noticeable percentage reporting high stress (4-5) and isolation.

Company Support: Satisfaction with remote work and company support for remote work correlates positively with work-life balance ratings.

**Preliminary Statistics**

Age: Median = 35 years; Range = 22-65 years.

Stress Level: Mean = 3.2; SD = 1.1.

Work Hours: Mean = 42 hours/week; SD = 10.5 hours.

**Relationships**

A preliminary scatter plot of “Work_Life_Balance_Rating” vs. “Stress_Level” indicates a negative correlation.

Grouping by “Industry” shows variations in satisfaction with remote work, suggesting industry-specific interventions may be necessary.

**Verifying Data Quality:**

**Completeness**

Critical Columns: Columns like “Stress_Level” and “Mental_Health_Condition” have minimal missing data less than 5%.

Demographic Fields: Slightly higher missing values almost 10% in “Region” and “Years_of_Experience.”

**Consistency**

The dataset appears consistent in categorical fields.

Some inconsistencies observed in numeric columns (“Hours_Worked_Per_Week” has a few entries exceeding 100.)

**Accuracy**

Self-reported data introduces potential biases, such as underreporting of stress levels or overreporting of productivity changes.

Verification with external datasets (if feasible) can strengthen reliability.

**Outliers**

Notable outliers in “Hours_Worked_Per_Week” and “Sleep_Quality” require further investigation to avoid skewing the analysis.

**Ethical Considerations**

Ensure anonymization of all identifiers.

Avoid aggregating data in ways that may indirectly expose individual identities, especially in small demographic groups.

**Decision on Data Usage**

Include all fields initially, focusing on filtering for relevant data.

Perform additional cleaning and transformations in the data preparation step to handle outliers and missing data effectively.

Exclude fields with excessive missing or unreliable data if no imputation method is appropriate.

**Conclusion**

The dataset offers a comprehensive basis for analysis of remote work's impact on mental health. Preliminary exploration confirms the feasibility of achieving the business and data-mining goals, with some considerations for handling missing data, outliers, and potential biases during subsequent steps.

**Task 4. Project plan**

**1. Data Setup:**

- Download and import the dataset (Daria, 0.5 hour)
- Initial exploration: data types, missing values (Anton,1h)
- Handle missing values (Anton, 1 h)
- Visualize key variable distributions (Elisabeth, 1.5 h)

**2. Exploratory Data Analysis:**

- **Univariate:**
- Explore Numeric Variables. Calculate summary statistics. (Elisabeth, 4 hours).
- Explore categorical variables. Analyze the frequency and proportion of different categories. (Daria, 3 h)
- **Bivariate:**
- Analyze relationships between Stress_Level and other numerical variables (Elisabeth, 4 h).
- Explore Job_Role vs. Stress_Level (Daria, 3 h).
- Investigate Company_Support impact on Work_Life_Balance (Anton, 3 h).
- **Multivariate:**
- Analyze how Mental_Health_Condition varies across different demographic groups (e.g., Gender, Age, Job_Role, Industry). Create crosstabs and stacked bar charts to visualize these relationships. (Elisabeth, 4 h)

**4. Machine Learning:**

- Select models to predict Stress_Level (All, 1 h).
- Train and evaluate models (Anton, Daria, Elisabeth, 4 h per person).
- Tune hyperparameters (Anton, Daria, Elisabeth, 2 h per person)).

**5. Clustering:**

- Perform k-means clustering (Anton, 6 h).
- Analyze and interpret clusters (Elisabeth, 1 h; Daria, 3 h).

**6. Visualization & Reporting:**

- Create visualizations (Anton, Daria, Elisabeth, 3 h per person).
- Write the final report (All, 3 hours).