**Resources:**

* Jupyter notebook, accessed via the remote desktop environment (RDE)
* The Mod 5 Fusion Day folder within Python
* The Python script file: **Module 5 Fusion Day.ipynb**
* Excel files:
  + Bike\_data\_2021\_part1.xlsx
  + Bike\_data\_2021\_part2.xlsx
  + Bike\_data\_new.xlsx

|  |
| --- |
| **Task 1: Collaborative discussion** |
| **Output:** Your discussion notes |
| **Time:** 10 mins |
| You’ll join a small group of your fellow analysts in a **virtual breakout room** to work together to answer the following question:   * Transport accounts for 23% of the world’s energy-related carbon dioxide (CO2) emissions. What role can data play in reducing this figure?   Source: [International Institute for Sustainable Development](https://www.iisd.org/articles/road-sustainable-transport)  **Instructions:**   1. Join the discussion 2. Add your thoughts to your group’s Jamboard 3. Type any discussion notes you’d like to take from this task in the space below |
| Type your discussion notes here: |

**About the dataset**

There are 14 variables included in the datasets; these are listed below. The ‘bikecount’ variable is the dependent variable. The dataset tracks specific weather conditions that may affect the dependent variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Name | Data type | Description | Levels |
| 1 | timestamp | Data-time | A unique date and time stamp | None |
| 2 | year | Categorical | The year 2021 | 2021 |
| 3 | season | Categorical | Meteorological seasons | Spring = Mar to May  Summer = June to Aug  Autumn = Sept to Nov  Winter = Dec to Feb |
| 4 | month | Categorical | The month of the year: January to December | 1 = January  2 = February  3 = March  4 = April  5 = May  6 = June  7 = July  8 = August  9 = September  10 = October  11 = November  12 = December |
| 5 | day | Categorical | The day of the week: Monday to Sunday | 1 = Monday  2 = Tuesday  3 = Wednesday  4 = Thursday  5 = Friday  6 = Saturday  7 = Sunday |
| 6 | hour | Categorical | Hours from 0 to 23 | 0–23 |
| 7 | isholiday | Categorical | Shows whether or not the day is a public holiday | 1 = holiday  0 = non-holiday |
| 8 | isweekend | Categorical | Shows whether the day is a weekend or a weekday | 1 = weekend  0 = weekday |
| 9 | weathercode | Categorical | Shows the day’s weather status | 1 = mostly clear, but some areas may have patches of fog/haze  2 = scattered clouds or few clouds  3 = Broken clouds  4 = Cloudy  7 = Rain or light rain shower  10 = Rain with thunderstorm  26 = Snowfall  94 = Freezing Fog |
| 10 | t1 | Numerical | Shows the real temperature in degrees Celsius | None |
| 11 | t2 | Numerical | Shows the ‘feels like’ temperature in degrees Celsius | None |
| 12 | humidity | Numerical | Shows the humidity as a percentage | None |
| 13 | windspeed | Numerical | Shows the wind speed in km/h | None |
| 14 | bikecount | Numerical | Shows the count of new bike shares | None |

 'Powered by TfL Open Data'

'Contains OS data © Crown copyright and database rights 2016' and Geomni UK Map data © and database rights [2019]

|  |
| --- |
| **Task 2: Inventory analysis** |
| **Output:** Your completed Jupyter notebook |
| **Time:** 3hr (excluding checkpoints) |
| Your client, Transport for London (TfL), wish to improve the overall inventory efficiency for their London cycle hire scheme. They’re implementing a new inventory management process and have recruited the help of Convvey, our fictional data agency.  TfL need to ensure their cycle inventory levels meet customer demand for the coming year. Your task is to answer TfL’s business question:  ‘**What is the predicted demand for the London cycle hire scheme for the next 12-month period?**’  To complete this task, follow the instructions set out in parts **a)** to **e)** below. |
| **a) Import and combine the data - 5 mins** |
| **Instructions:**   1. Access **Jupyter Notebook** using the remote desktop environment (RDE) 2. Set the Mod 5 Fusion Day folder as your default folder in Jupyter Notebook 3. Open the Python script file **Module 5 Fusion Day.ipynb** 4. Install and load the Python libraries 5. From the **Mod 5 Fusion Day folder**, import the Excel files listed below into the Jupyter Notebook working environment:    * Bike\_data\_2021\_part1.xlsx    * Bike\_data\_2021\_part2.xlsx 6. Combine the two imported datasets listed above to create one dataset for analysis    1. Save the newly combined dataset with the name ‘df’ |
| **b) Data preparation - 1hr** |
| **Instructions:**   1. Carry out **data quality checks** to detect issues within the dataset    1. View sample of the data    2. Identify the shape of the data    3. Identify if there are any duplicate rows    4. Identify if there is any missing data 2. Carry out **data cleansing**     1. Convert variables to appropriate datatypes    2. Remove duplicate rows    3. Interpolate missing values using the median   **Hint:** After carrying out data cleansing, your dataset should have 14 variables and 8,675 entries. |
| **Checkpoint 1 – 15 mins**  (Optional)  Check-in with your tutor and fellow analysts to discuss section a) and b). |
| **c) Exploratory data analysis and visualisation – 1hr** |
| Perform exploratory data analysis (EDA) using the following three techniques:   * Aggregation * Distribution * Correlation   For each technique, make a note of your findings in the box below. To complete your EDA, you need to visualise your data. What interesting findings can you uncover?    **Instructions:**   1. Perform EDA using appropriate **aggregation** functions and tools.    1. Calculate the **total** number of bike shares    2. Calculate the **average** number of bike shares 2. Perform EDA using appropriate groupings and aggregate functions (**distribution**. Calculate the total average bike shares for the following:    1. **Season**. Which season has the highest average value?    2. **Month**. Which month has the highest average value?    3. **Day of the week.** Which day has the highest average value?    4. **Day of the week for the most popular month (2.2)** with the **highest** **average value**    5. Calculate all the descriptive statistics for all numerical variables.    6. What type of distribution does the bike share have – normal or skewed? 3. Continue your EDA working with numerical variables    1. Create a correlation matrix for the numerical variables only    2. In the Python script, make a note of **which variables are strongly correlated** 4. Complete your EDA by creating three **visualisations** of the data – a jointplot, a lineplot and a pairgrid. To speed up this process, we have provided the code blocks you need below    1. Use the code block below to produce a **jointplot**:   sns.jointplot(data=df, x="bikecount", y="t1", hue="season")  plt.xlabel("Bike Count")  plt.ylabel("Temperature")  plt.show()   * 1. Use the code block below to produce a **lineplot**:   sns.lineplot(data=df, x="month", y="bikecount", hue="isweekend")  plt.title("Bike share analysis")  plt.xlabel("Month")  plt.ylabel("Bike Count")  plt.show()   * 1. Use the code below to produce a **pairgrid**:   g = sns.PairGrid(data=df, hue="season", vars=["t1","t2","humidity","windspeed","bikecount"])  g.map\_diag(plt.hist)  g.map\_offdiag(plt.scatter)  g.add\_legend()   * 1. Study your visualisations. Are there any interesting behaviours in the data? Make a note of any interesting findings in the Python script. |
| **Checkpoint 2 – 15mins**  (Optional)  Check-in with your tutor and fellow analysts to discuss section c). |
| **d) Predictive modelling (regression) - 50 mins** |
| **Instructions:**   1. **Pre-processing**    1. Remove the ‘timestamp’ and ‘year’ variables from the dataset    2. Encode all categorical data   **Hint:**  After carrying out pre-processing, the dataset should have the shape (8675, 56).   1. **Train/ test split (70%/30%)**    1. Split the df dataset into x and y datasets    2. Perform 70/30 random split on x and y datasets using random state = 1 2. Build a **linear regression model** using the **training datasets**    1. Save the regression function ‘LinearRegression()’ into a container called ‘model’    2. Fit the regression into the training data 3. **Evaluate** your model of the **training dataset**    1. Identify regression intercept       1. Reformat intercept into a dataframe    2. Identify regression coefficients       1. Reformat coefficients into a dataframe 4. **Run** your model on the **testing dataset** 5. Calculate R squared AND RMSE (root mean squared error), and then intercept the results |
| **e) Interpret your results – 5 mins** |
| Interpret the results of your model.  Identify any key findings and develop a narrative to summarise your findings. Type your summary in the Python script.  **Hint:** When writing your summary, consider your answers to the following questions:   * Which aspects of the results stand out the most? * Are there any adjustments that you need to make to the model, to make it more effective? |

|  |
| --- |
| **Task 3: Model predictions** |
| **Output:**   * A csv file showing the predicted demand for the London cycle hire scheme for the next 12-month period |
| **Time: 1hr** |
| You’ll need the ‘bike\_data\_new.xlsx file’ to complete this task.  Use your linear regression model and the data provided in the file above to answer TfL’s business question:  ‘**What is the predicted demand for the London cycle hire scheme for the next 12-month period?**’  **Instructions:**   1. From the Mod 5 Fusion Day folder, import the dataset ‘**bike\_data\_new.xlsx**’ 2. Perform data quality checks 3. Perform data cleaning    1. Remove duplicates    2. Remove missing data    3. Make a copy of the dataset    4. Convert variables to the appropriate data types 4. Pre-processing    1. Remove the timestamp and year columns    2. Encode all categorical data 5. Prediction    1. Use your model to predict bike shares    2. Combine predicted bike shares from 5.1 with the new data (bike\_data\_new.xlsx)       1. Reformat predictions into a datafram called ‘dfr’       2. Create a new dataframe that joins dfr to the ‘dfp\_copy’ 6. Export the predicted results as a csv file to save your predictions |

|  |
| --- |
| **Regather for a team debrief at 16.00–16.30** |
| **Time:** 30 mins |
| Join your tutor and fellow analysts to discuss the outputs of the day.  In this session, your tutor will share worked examples of the tasks you have completed. You’ll be invited to share what worked well for you and what didn’t work so well. You’ll also discuss what you might do differently next time when faced with a similar task at work.  This is an opportunity for you to share your own insights and also learn from others. |