This document contains privileged information as property of GfK SE.
Please note that unauthorized copying, disclosure, or distribution of the material in this document is not permitted.

General Set-Up

Please commit all code & results in a public Git repository (e.g., GitHub) and share the link to the repository with us. Please also make sure to not mention “GfK SE” anywhere in the repository (e.g., repository name, README, description etc.).

If you have any questions or something in the task description is unclear, please feel free to contact us anytime.

1 SQL

Tasks:

1. Setup a pod or isolated container comprising a SQL Database
2. Import testset_B.tsv into the SQL DB
3. Create a Python script that connects to the database, calculates the following KPIs and stores each of them in a separate CSV file:
   a. Ranks based on column “Price”, grouped by column “brand”
   b. Minimum and maximum of column “HDD_GB”
   c. Median of column “GHz”, grouped by column “RAM_GB”
4. Describe all your steps in a README file and commit all code & results to a Git repository.

2 ML Tasks

Please choose **one** of the following two tasks. You only need to work on and present the results for **one** task, depending on your preference.

2.1 Product Category Classification

Dataset C: testset_C.csv

Dataset description:
Dataset C is supposed to contain records with article texts and their corresponding product group. The following information is known about the columns:

Column	**Info**
id	A unique record identifier
product group	Product category
main_text	A describing text about the article
add_text	An additional describing text about the article
manufacturer	The manufacturer belonging to the article

Tasks:

1. Create a machine learning model that predicts the product category based on appropriate columns.
2. Present the result in a vivid way and explain your model from a statistical point of view (e.g., in a Jupyter notebook).
3. Create an HTTP REST-API on top of your model which obtains an article text and returns the product category.
4. Add a README and at least one unit test for your endpoint. Also, propose some additional steps to make your code “production-ready” (you don’t actually need to implement all of them).
5. Commit all code & results to a Git repository.

2.2 House Price Prediction

Dataset:

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Tasks:

1. Create a model with the ability to predict house prices (Y), based on square feet as input parameter (X)
2. Present the result in a vivid way and explain your model from a statistical point of view (e.g., in a Jupyter notebook).
3. Create an HTTP REST-API on top of your model that takes X as parameter for the request and responds with prediction Y
4. Add a README and at least one unit test for your endpoint. Also, propose some additional steps to make your code “production-ready” (you don’t actually need to implement all of them).
5. Commit all code & results to a Git repository.