

Data analysis using openAI: *testing their new capabilities focusing on data science*

Jorge Guerra Pires^{1*}

Abstract: This are the Supplementary Material for the main paper. On this document, you have the full conversation with openAI coder interpreter. See the main paper for more details.
: openAI — data science — statistical inference — python — open source — diabetes

¹ Founder at IdeaCoding Lab and JovemPesquisador.com

1. Understanding apparent temperature vs. temperature

1.1 Details

It was uploaded a spreadsheet directly. The spreadsheet from Kaggle, has information about temperature and apparent temperature, plus a set of other measurements.

You can find it here the spreadsheet used on the analysis. It contains several columns, more than 6.000 samples for the analysis. It is a massive number of samples. It is worth-noting that we are not analyzing a small spreadsheet, it is huge one. This is impressive that this tool from openAI can make such a analysis so fast and effortless.

We have used their new capability called *Assistant as Code Interpreter*. We have used the assistant on Playground.

The instruction we gave, in addition to attach the CSV file of the samples.

Attached is a set of observations of the Apparent Temperature (C) and the Temperature (C) in CSV, also other measurements. Your job is to analyze this dataset. and answer my questions based on this analysis.

1.2 Conversation

User: I want to understand better the relation between Apparent Temperature and Temperature.

Whether bot: Sure, I can help you with that. Let's start by loading the dataset so we can take a look at the contents and then proceed with the analysis.

Whether bot: The dataset contains the following columns:

- Formatted Date
- Summary
- Precip Type
- Temperature (C)
- Apparent Temperature (C)
- Humidity
- Wind Speed (km/h)
- Wind Bearing (degrees)
- Visibility (km)
- Loud Cover
- Pressure (millibars)
- Daily Summary

The main columns of interest for analyzing the relationship between Apparent Temperature and Temperature are "Temperature (C)" and "Apparent Temperature (C)".

Now, I will proceed with performing some basic statistical analysis and visualizations to better understand the relationship between these two variables.

Obs. what is interesting, which we are not replicating here: it is actually calling well-established libraries in Python, you can see on the logs before each response. Python has a rich environment for data science. They provide other APIs that can be used to adapt that for your own program/library.

Whether bot: Above [fig.1] is a scatter plot showing the relationship between Temperature and Apparent Temperature. Each point on the

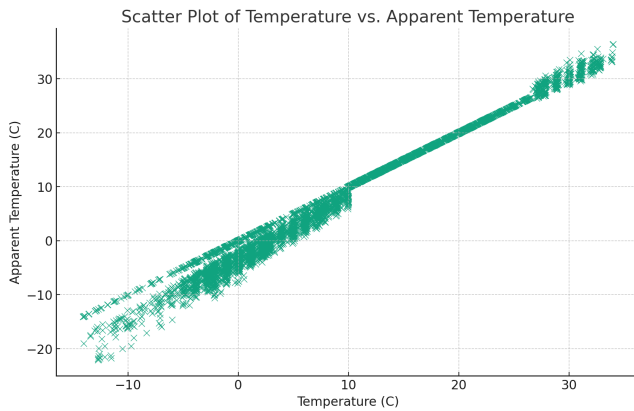


Figure 1. Temperature vs. Apparent Temperature. Source: plotted by openAI.

plot represents a specific date and time, with the x-coordinate representing the Temperature and the y-coordinate representing the Apparent Temperature. [see that I did the same when I have studied this dataset manually]

As a next step, I can calculate the correlation coefficient between these two variables to quantify the strength and direction of the linear relationship between them. Additionally, I can provide some statistical summary to further analyze their relationship.

Whether bot: The correlation coefficient between Temperature and Apparent Temperature is 0.993, indicating a very strong positive linear relationship between these two variables. This suggests that as Temperature increases, Apparent Temperature also tends to increase, and vice versa.

Here is a statistical summary of the Temperature and Apparent Temperature:

- Mean Temperature: 10.22°C
- Mean Apparent Temperature: 8.96°C
- Standard Deviation of Temperature: 10.57°C
- Standard Deviation of Apparent Temperature: 11.92°C
- The range of values for both Temperature and Apparent Temperature

If you have further questions or need additional analysis, feel free to ask!

Obs. see that a single question created all this sequence of responses, looking for an way best answer the question using the given dataset. This is a new level of intelligence: the machines now can decide, use tools, for achieving a goal.

User: If we look at the tails of the plotted graph [fig. 1], you can see it is wider (the data is noisier), and the center is narrow (the correlation is stronger). Is there any explanation for that?

The observation you made regarding the width of the tails of the plotted graph and the strength of correlation is insightful. This pattern is often observed in datasets where the relationship between two variables is influenced by external factors. In the context of weather data, such as the one we are analyzing, this pattern could be attributed to the following factors:

1. **Weather Variability:** In extreme weather conditions, the difference between the temperature and the apparent temperature can be more pronounced, leading to a wider spread of data points in the tails of the plot. For example, in very cold or very hot weather, the discrepancy between the actual temperature and the perceived (apparent) temperature can be significant.
2. **Humidity and Wind:** The apparent temperature takes into account factors such as humidity and wind, which can affect how the temperature is perceived by humans. In conditions of high humidity or strong winds, the apparent temperature can deviate more from the actual temperature, leading to a wider spread of data points.
3. **Measurement Error:** Variability in the measurement or estimation of both temperature and apparent temperature can also contribute to the noise in the data. Factors such as instrument accuracy, positioning, and local environmental variations can introduce variability in the recorded values.

The combination of these factors can lead to the observed pattern, with a wider spread of data points in the tails and a narrower center, reflecting the noisier nature of the data in extreme conditions and the stronger correlation in more moderate conditions.

If you would like to explore these factors further or have any other questions, please feel free to ask!

2. Understanding diabetic signals

We on this part, we are going to explore the same dataset from [1]. On their scenario, they did all the analysis manually. We are going to compare with code interpreter, and see the advantages and disadvantages of this new openAI's tool.

The questions we did here mirrors the ones we did at [1] for building the respective model.

The complete CSV file is here.

2.0.1 Conversation

User: Please, describe the dataset

DiabetesBot: The dataset contains the following columns:

1. gender: Gender of the patient
2. age: Age of the patient
3. hypertension: Whether the patient has hypertension (1 for yes, 0 for no)
4. heart_disease: Whether the patient has heart disease (1 for yes, 0 for no)
5. smoking_history: Smoking history of the patient
6. bmi: Body Mass Index (BMI) of the patient
7. HbA1c_level: HbA1c level of the patient
8. blood_glucose_level: Blood glucose level of the patient
9. diabetes: Whether the patient has diabetes (1 for yes, 0 for no)

This dataset contains information about diabetic patients, including their demographics, medical history, and health indicators.

Obs. what is interesting about this part: I have not provided any legend, it guessed what each column meant.

User: Can you make a hypothesis test whether diabetes is more prevalent on man or female?

DiabetesBot: To test whether diabetes is more prevalent in men or women, we can conduct a hypothesis test to compare the proportion of diabetic patients between the two genders. We will use a two-sample hypothesis test for proportions to determine if there is a significant difference.

The null hypothesis (H_0) is that there is no difference in the proportion of diabetic patients between men and women. The alternative hypothesis (H_1) is that there is a difference in the proportion of diabetic patients between men and women.

Let's perform the hypothesis test and determine whether there is a significant difference in the prevalence of diabetes between men and women.

DiabetesBot: The p-value for the two-sample proportion hypothesis test comparing the prevalence of diabetes between men and women is

approximately $1.24e-32$, which is extremely low. This indicates strong evidence against the null hypothesis (H_0) that there is no difference in the proportion of diabetic patients between men and women.

Therefore, we reject the null hypothesis in favor of the alternative hypothesis (H_1) that there is a difference in the prevalence of diabetes between men and women. This suggests that the prevalence of diabetes differs significantly between genders in the given dataset.

Obs. See that it performed an hypothesis test, without human guidance.

User: Which one is most likely to have diabetes: men or women?

DiabetesBot: Based on the results of the hypothesis test, we have found that there is a significant difference in the prevalence of diabetes between men and women. To determine which gender is most likely to have diabetes, we can compare the proportions of diabetic patients for each gender and identify which gender has a higher proportion of diabetic patients.

Let's compare the proportions of diabetic patients for men and women to determine which gender is most likely to have diabetes in the given dataset.

DiabetesBot: The proportion of diabetic patients is approximately 7.6% for women and 9.7% for men in the given dataset. Therefore, based on these proportions, men are more likely to have diabetes compared to women.

User: I already know that "The proportion of diabetic patients is approximately 7.6% for women and 9.7% for men in the given dataset.", as you have calculated. Can you also calculate a confidence interval for those proportions?

DiabetesBot: The 95% confidence intervals for the proportions of diabetic patients in the given dataset are approximately:

- For women: 7.41% to 7.84%
- For men: 9.47% to 10.04%

These intervals provide a range of values within which we can be 95% confident that the true proportion of diabetic patients lies for each gender.

User: Is there any possible explanation for that?

DiabetesBot: There are several possible explanations for the higher prevalence of diabetes in men compared to women in the given dataset. Here are some potential factors that could contribute to this observation:

1. **Differences in Risk Factors:** Men and women may have different risk factors for diabetes. For example, men may be more likely to have higher levels of central obesity or visceral fat, which are associated with an increased risk of diabetes.
2. **Biological Differences:** Biological differences between men and women, such as hormone levels and metabolism, may contribute to varying susceptibility to diabetes.
3. **Behavioral Factors:** Differences in lifestyle and behaviors, such as dietary habits, physical activity levels, and smoking history, could play a role in the higher prevalence of diabetes in men.
4. **Healthcare Utilization:** Variations in healthcare utilization and access to preventive care between men and women could impact the diagnosis and management of diabetes.
5. **Genetic and Environmental Factors:** Genetic predisposition to diabetes and environmental factors may differ between men and women, contributing to variations in diabetes prevalence.

It's important to note that these are potential explanations, and further research and analysis would be needed to fully understand the underlying reasons for the observed difference in diabetes prevalence between men and women in the dataset.

User: Given the features that can predict diabetes. Which ones have the highest correlation on predicting diabetes?

DiabetesBot: To identify the features that have the highest correlation with predicting diabetes, we can calculate the correlation coefficients between each feature and the "diabetes" column in the dataset. Features with higher absolute correlation coefficients are more strongly associated with predicting diabetes.

Let's calculate the correlation coefficients between each feature and the "diabetes" column and identify the features that have the highest correlations with predicting diabetes.

DiabetesBot: The correlation coefficients between the features and the "diabetes" column are as follows:

1. Blood Glucose Level: 0.4196
2. HbA1c Level: 0.4007
3. Age: 0.2580
4. BMI (Body Mass Index): 0.2144
5. Hypertension: 0.1978
6. Heart Disease: 0.1717

The features with the highest correlations (in absolute value) with predicting diabetes are "Blood Glucose Level" and "HbA1c Level," indicating that these two features are strongly associated with predicting diabetes in the dataset. These results suggest that blood glucose level and HbA1c level may be important indicators for predicting the presence of diabetes in the given dataset.

User: If someone has high blood glucose levels for days; HbA1c is measured in months. Would it be enough to predict diabetes?

DiabetesBot: While high blood glucose levels and HbA1c levels are important indicators for predicting diabetes, it's crucial to consider the temporal aspect and the limitations of relying solely on individual measurements.

1. **Blood Glucose Levels:** High blood glucose levels are a significant indicator of diabetes. However, blood glucose levels can fluctuate throughout the day, affected by factors such as diet, physical activity, and stress. Therefore, a single measurement of blood glucose may not provide a comprehensive picture of an individual's diabetic status. Continuous monitoring of blood glucose levels over several days or weeks can provide more meaningful insights into an individual's glycemic control and the presence of diabetes.
2. **HbA1c Levels:** HbA1c levels reflect average blood glucose levels over the past 2-3 months. While HbA1c provides a longer-term view of glycemic control, it may not capture short-term fluctuations in blood glucose levels or sudden changes in diabetic status. Additionally, individual variation in red blood cell turnover and other factors can impact HbA1c levels.

Given the temporal aspects of blood glucose and HbA1c measurements, it is important to consider the following:

- **Longitudinal Monitoring:** Longitudinal monitoring of blood glucose levels and regular HbA1c testing over an extended period is essential for assessing an individual's diabetic status and the effectiveness of diabetes management.
- **Clinical Evaluation:** Diabetes diagnosis and management should involve a comprehensive clinical evaluation, including medical history, physical examination, and possibly additional tests and assessments.
- **Holistic Approach:** While blood glucose and HbA1c levels are important, a comprehensive approach to diabetes diagnosis and management should consider multiple factors, including patient history, symptoms, risk factors, and potential complications.

In summary, while high blood glucose and elevated HbA1c levels are significant indicators for predicting diabetes, a holistic and longitudinal approach, including clinical evaluation and monitoring, is crucial for accurate diagnosis and management of diabetes.

DiabetesBot: As we've discussed, relying solely on a single measurement of blood glucose or HbA1c may not be sufficient to predict diabetes. Instead, a comprehensive approach involving longitudinal monitoring, clinical evaluation, and consideration of multiple factors is important for accurate diagnosis and management. If you have any further questions or if there's anything else I can assist you with, feel free to let me know!

User: It seems some medical conditions and even some treatment may increase glucose levels, without necessarily being diabetes. Therefore, those measure alone can be masked, misleading.

DiabetesBot: Absolutely, that's an important point. Various medical conditions, medications, and treatments can impact blood glucose levels, leading to elevated readings that do not necessarily indicate the presence of diabetes. Conditions such as stress, infections, certain medications (e.g., corticosteroids), hormonal disorders, and pancreatic diseases can all affect blood glucose levels. Additionally, specific treatments, such as corticosteroid therapy or certain medications, may also lead to elevated blood glucose levels.

Therefore, relying solely on blood glucose levels or HbA1c measurements for diabetes diagnosis without considering the broader clinical context can indeed be misleading. It underscores the

importance of a comprehensive approach to diabetes diagnosis, including a thorough medical history, clinical evaluation, and consideration of potential contributing factors to elevated blood glucose levels.

Understanding the nuanced relationship between medical conditions, treatments, and blood glucose levels is crucial for accurate interpretation and assessment. This underscores the need for healthcare professionals to consider the full clinical picture and to conduct appropriate follow-up and additional evaluation when interpreting blood glucose and HbA1c measurements.

Your insight highlights the complexities involved in interpreting these measurements and the significance of a holistic approach to medical assessment. If you'd like further information on this topic or have additional questions, feel free to ask!

3. Should we buy a house in Boston?

We are going to explore an well-known dataset in machine learning: the Boston House dataset [2, 3, 4] . This dataset provides prices for houses in Boston with local information such as crime rate. It can be used for predicting the value of a house on that region.

"If we were the only ones in the world with access to this info, we could be the best Boston real-estate investors in 1978! Unless, somehow, someone were able to build an even more accurate estimate . . ." [highlight added] [3]

User: Describe the dataset

BrokerBotThe dataset contains 506 entries and 14 columns. Here are the columns and their data types:

1. CRIM: per capita crime rate by town (float)
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft. (float)
3. INDUS: proportion of non-retail business acres per town (float)
4. CHAS: Charles River dummy variable (int)
5. NOX: nitric oxides concentration (parts per 10 million) (float)
6. RM: average number of rooms per dwelling (float)
7. AGE: proportion of owner-occupied units built prior to 1940 (float)

8. DIS: weighted distances to five Boston employment centers (float)
9. RAD: index of accessibility to radial highways (int)
10. TAX: full-value property-tax rate per \$10,000 (float)
11. PTRATIO: pupil-teacher ratio by town (float)
12. B: $1000(Bk0.63)^2$ where Bk is the proportion of blacks by town (float)
13. LSTAT: % lower status of the population (float)
14. MEDV: Median value of owner-occupied homes in \$1000s (float)

The data types are either float or integer, and there are no missing values in the dataset.

User: Let me know the correlation by the price of the houses and crime rate

BrokerBot: The correlation coefficient between the median value of owner-occupied homes (MEDV) and the per capita crime rate by town (CRIM) is approximately -0.3883. This indicates a moderate negative correlation between the two variables, suggesting that areas with higher crime rates tend to have lower median house values.

User: What about number of teacher per student and prices of the houses?

BrokerBot: The correlation coefficient between the median value of owner-occupied homes (MEDV) and the pupil-teacher ratio by town (PTRATIO) is approximately -0.5078. This indicates a moderate negative correlation, suggesting that areas with higher pupil-teacher ratios tend to have lower median house values.

3.1 Can we make any hypothesis test for proving that crime rate will influence on the house pricing

User: Could you test the hypothesis that high crime rate regions will have lower market value?

BrokerBot: To test the hypothesis that high crime rate regions will have lower market value, we can conduct a simple correlation analysis between the "CRIM" (per capita crime rate) and "MEDV" (median value of owner-occupied homes) attributes in the dataset. This will allow us to determine if there is a relationship between crime rate and home values.

First, I will load the dataset and take a look at the first few rows to understand its structure.

BrokerBot: The dataset has been loaded and contains the following columns:

- CRIM: Per capita crime rate by town
- ZN: Proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: Proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: Nitric oxides concentration (parts per 10 million)
- RM: Average number of rooms per dwelling
- AGE: Proportion of owner-occupied units built prior to 1940
- DIS: Weighted distances to five Boston employment centers
- RAD: Index of accessibility to radial highways
- TAX: Full-value property-tax rate per \$10,000
- PTRATIO: Pupil-teacher ratio by town
- B: $1000(Bk0.63)^2$ where Bk is the proportion of blacks by town
- LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes in \$1000s

I will now calculate the correlation between the "CRIM" and "MEDV" attributes to examine the relationship between crime rate and home values.

BrokerBot: The correlation coefficient between the "CRIM" (per capita crime rate) and "MEDV" (median value of owner-occupied homes) attributes is approximately -0.39. This negative correlation suggests that there is a moderate inverse relationship between crime rate and home values. In other words, as the crime rate increases, the median home value tends to decrease.

However, correlation does not imply causation, so further analysis and possibly modeling would be needed to establish a causal relationship.

Would you like to perform any additional analysis or explore this relationship further?

User: Could you perform a hypothesis test?

BrokerBot: Certainly! To perform a hypothesis test, we can use a simple linear regression model to test the relationship between the crime rate and median home values. The null hypothesis for the test would be that there is no relationship

between the crime rate and median home values, while the alternative hypothesis would be that there is a significant relationship.

I will conduct a hypothesis test using a simple linear regression model and examine the p-value associated with the "CRIM" coefficient to determine if the relationship is statistically significant.

BrokerBot: The hypothesis test using a simple linear regression model yields the following results:

- The p-value associated with the coefficient of "CRIM" is very small (approximately 0.000), indicating that the coefficient is statistically significant.
- The null hypothesis that there is no relationship between the crime rate and median home values is rejected in favor of the alternative hypothesis that there is a significant relationship.

The coefficient of "CRIM" is estimated to be approximately -0.415, indicating that for every one unit increase in per capita crime rate, the median home value is expected to decrease by approximately \$415.

Based on these results, we can conclude that there is a statistically significant relationship between the crime rate and median home values in the dataset.

Is there anything else you would like to explore or analyze related to this hypothesis test?

Obs. Even though the information provided is valuable, it did not performed what I wanted. Let's try being more detailed.

User: Can you test the hypothesis that the crime rate is higher than the average for regions of high crime rate? For that, calculate the average crime rate, and then separate the dataset set into above the average and lower than the average and test the hypothesis that the house pricing are lower at the group with crime rate above the average.

BrokerBot: Certainly! To test the hypothesis that the crime rate is higher than the average for regions of high crime rate, we can follow these steps:

1. Calculate the average crime rate for all regions.
2. Split the dataset into two groups: one with crime rates above the average and one with crime rates below the average.

3. Conduct a hypothesis test to compare the median home values between the two groups.

Let's proceed with these steps to test the hypothesis. I will start by calculating the average crime rate for all regions in the dataset.

Obs. interesting enough, it was able to break down my request into sub-tasks.

BrokerBot: The average crime rate for all regions in the dataset is approximately 3.61 per capita.

Now, I will split the dataset into two groups: one with crime rates above the average and one with crime rates below the average. Then, I will conduct a hypothesis test to compare the median home values between the two groups.

BrokerBot: The hypothesis test comparing the median home values between regions with crime rates above the average and regions with crime rates below the average yields the following results:

- The t-statistic is approximately -9.70.
- The p-value is approximately 1.16×10^{-18} , indicating that it is extremely small.

Based on these results, we reject the null hypothesis and conclude that there is a statistically significant difference in median home values between regions with crime rates above the average and regions with crime rates below the average. Specifically, the median home values are lower in regions with crime rates above the average.

Is there anything else you would like to explore or analyze?

Obs. on the case of the apparent temperature vs. temperature I was able to create a neural network and make predictions. Now, I was just able to create the neural model, I was unable to make prediction. Therefore, the current environment they have seem not totally reliable for machine learning.

One part of the conversation for creating a neural model:

BrokerBot: It looks like I'm unable to run the code to create and train a neural network model at the moment because the required deep learning libraries such as TensorFlow and Keras are not available in this environment.

If you have the capability to run the code in a local environment where TensorFlow and Keras are available, I can provide you with the necessary code and instructions. Alternatively, if there is anything else I can assist you with, please feel free to let me know.

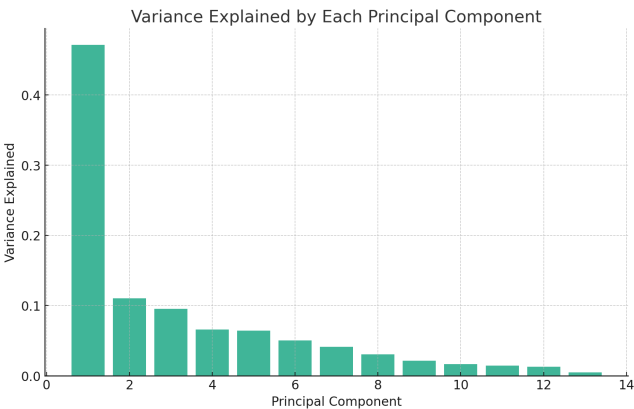


Figure 2. Principal Component Analysis for all features

	Feature	PC1_Weight
0	CRIM	0.250951
1	ZN	-0.256315
2	INDUS	0.346672
3	CHAS	0.005042
4	NOX	0.342852
5	RM	-0.189243
6	AGE	0.313671
7	DIS	-0.321544
8	RAD	0.319793
9	TAX	0.338469
10	PTRATIO	0.204942
11	B	-0.202973
12	LSTAT	0.309760

Figure 3. Enter Caption

What can we learn from a Principal Component analysis (PCA)

BrokerBot: Above is the bar plot showing the variance explained by each principal component. As we can see, the first few principal components capture the majority of the variance in the data, while the later components capture less and less variance.

Is there anything else you would like to explore or analyze in relation to the principal components, or any other aspect of the dataset you'd like to delve into?

BrokerBot: The table above shows the feature weights for the first principal component (PC1). These weights indicate the contribution of each

original feature to the first principal component. Positive weights indicate a positive correlation with PC1, while negative weights indicate a negative correlation.

Interpreting the feature weights:

- Features with larger absolute weights contribute more to the first principal component.
- Features with positive weights (e.g., INDUS, NOX, and TAX) have a positive influence on the first principal component, while features with negative weights (e.g., ZN, RM, and DIS) have a negative influence.
- The magnitude of the weights indicates the strength of the influence of each feature on PC1.

In summary, the first principal component (PC1) is a weighted sum of the original features, with each feature contributing differently based on its weight. It can be used to understand the most significant patterns or relationships present in the original features.

References

1 PIRES, J. G. Machine learning in medicine using javascript: building web apps using tensorflow.js for interpreting biomedical datasets. *medRxiv*, Cold Spring Harbor Laboratory Press, 2023. Disponível em: <<https://www.medrxiv.org/content/early/2023/07/09/2023.06.21.23291717>>.

2 PIRES, J. G. *Discussing a multiple regression model*. 2022. Disponível em: <<https://medium.com/computational-thinking-how-computers-think-decide/discussing-a-multiple-regression-model-c1f45ce52211>>.

3 NIELSEN, S. C. E. D.; BILESCHI, S. M.

4 PIRES, J. G. *Computational Thinking: How computers think, decide and learn, when human limits start and computers champ, vol. 1*. 2022. Disponível em: <https://books.google.com/books/about/Computational_Thinking_How_computers_thi.html?id=WfV6EAAAQBAJ>.