



# Building ResearchLLM

Automated statistical research and interpretation

## ResearchLLM

Ask a question! 😊

Submit

### Data Overview (Prompt)

I am researching the relationship between income and sociodemographic census info. I have imported Pandas as `pd`, Numpy as `np`, `scipy`, and `sklearn`, and have a dataframe called `df` loaded into Python. `df` contains the following variables and variable types:

Column Name: age  
Column Type: int64

Column Name: workclass  
Column Type: object  
Sample Values: ['Federal-gov', 'Private', 'Self-emp-inc', 'Local-gov', '?', 'Never-worked', 'State-gov', 'Self-emp-not-inc', 'Without-pay']

Column Name: fnlwgt  
Column Type: int64

Column Name: education  
Column Type: object  
Sample Values: ['10th', '9th', '5th-6th', 'Doctorate', 'Masters', 'Prof-school', 'HS-grad', 'Preschool', 'Some-college', 'Bachelors']

Column Name: education.num  
Column Type: int64

Column Name: marital.status  
Column Type: object  
Sample Values: ['Widowed', 'Divorced', 'Married-civ-spouse', 'Married-spouse-absent', 'Never-married', 'Separated', 'Married-AF-spouse']

Column Name: occupation  
Column Type: object  
Sample Values: ['Tech-support', 'Handlers-cleaners', 'Sales', 'Protective-serv', 'Armed-Forces', 'Exec-managerial', 'Craft-

## Analysis Output

Can you build a model to predict likelihood of being a high earner based on age?

Based on the outputs:

Coefficient for age variable: 0.03940502633208258

This indicates that for every 1 year increase in age, the log odds of being a high earner increases by 0.039 units. Since the log odds ratio is positive, this means that older individuals are more likely to be high earners, controlling for other factors.

Intercept: -2.7381512925135643

The negative intercept indicates that when age is 0, the predicted log odds of being a high earner is -2.738. In other words, very young individuals (age 0) are unlikely to be high earners according to the model.

Predicted likelihood of being a high earner (income > 50K) for age 35 : [0]  
For an individual aged 35 years old, the predicted likelihood of being a high earner is 0 according to the logistic regression model.

In summary, the model shows that increasing age has a small positive effect on the odds of being a high earner, but that an individual aged 35 years old still only has a 0% predicted likelihood of earning above 50K based on age alone. Additional sociodemographic variables would likely be needed to make an accurate prediction.

Does this help interpret the results? Let me know if you need any clarification or have additional questions.

[Show Code](#) [Show Python Output](#)



PhaseLLM dev tools:

<https://phasellm.com/>

PhaseLLM is open source:

<https://github.com/wgryc/phasellm>

ResearchLLM is part of PhaseLLM:

<https://phasellm.com/researchllm>

## What do I want?

An LLM to analyze my data and do research for me.

I don't want to share the data, though!

(Eventually, perform the entire research process for me.)



## ResearchLLM

Ask a question! 😊

Submit

### Data Overview (Prompt)

I am researching the relationship between income and sociodemographic census info. I have imported Pandas as `pd`, Numpy as `np`, `scipy`, and `sklearn`, and have a dataframe called `df` loaded into Python. `df` contains the following variables and variable types:

Column Name: age  
Column Type: int64

Column Name: workclass  
Column Type: object  
Sample Values: ['Federal-gov', 'Private', 'Self-emp-inc', 'Local-gov', '?', 'Never-worked', 'State-gov', 'Self-emp-not-inc', 'Without-pay']

Column Name: fnlwgt  
Column Type: int64

Column Name: education  
Column Type: object  
Sample Values: ['10th', '9th', '5th-6th', 'Doctorate', 'Masters', 'Prof-school', 'HS-grad', 'Preschool', 'Some-college', 'Bachelors']

Column Name: education.num  
Column Type: int64

Column Name: marital.status  
Column Type: object  
Sample Values: ['Widowed', 'Divorced', 'Married-civ-spouse', 'Married-spouse-absent', 'Never-married', 'Separated', 'Married-AF-spouse']

Column Name: occupation  
Column Type: object  
Sample Values: ['Tech-support', 'Handlers-cleaners', 'Sales', 'Protective-serv', 'Armed-Forces', 'Exec-managerial', 'Craft-

## Analysis Output

Can you build a model to predict likelihood of being a high earner based on age?

Based on the outputs:

Coefficient for age variable: 0.03940502633208258

This indicates that for every 1 year increase in age, the log odds of being a high earner increases by 0.039 units. Since the log odds ratio is positive, this means that older individuals are more likely to be high earners, controlling for other factors.

Intercept: -2.7381512925135643

The negative intercept indicates that when age is 0, the predicted log odds of being a high earner is -2.738. In other words, very young individuals (age 0) are unlikely to be high earners according to the model.

Predicted likelihood of being a high earner (income > 50K) for age 35 : [0]  
For an individual aged 35 years old, the predicted likelihood of being a high earner is 0 according to the logistic regression model.

In summary, the model shows that increasing age has a small positive effect on the odds of being a high earner, but that an individual aged 35 years old still only has a 0% predicted likelihood of earning above 50K based on age alone. Additional sociodemographic variables would likely be needed to make an accurate prediction.

Does this help interpret the results? Let me know if you need any clarification or have additional questions.

[Show Code](#) [Show Python Output](#)



## Startup Prompt

You are a data science helper and will be working with me to build a model to explore a data set. You do not need to provide qualifiers like "As an AI model" because I know you are such a model. I want you to be as productive and concise as possible with me.

If you are generating code in a response, please limit your code generation to ONE (1) code block. If this means you need to add additional comments in the code, this is perfectly fine. Precede every code block you generate with "`|--START PYTHON CODE--|`" and end each code block with "`|--END PYTHON CODE--|`".

Do you understand? Please simply write "yes" if you do, and "no" with followup questions if you do not.



```
Python 3.10.6 (main, Mar 10 2023, 10:55:28) [GCC 11.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> from frontend import *
>>> run()
Running ResearchLLM with ClaudeWrapper(model=claude-v1)
Understood? yes (attempt #1)
  * Serving Flask app 'frontend'
  * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment
  Use a production WSGI server instead.
  * Running on http://127.0.0.1:5000
Press CTRL+C to quit
```

## ResearchLLM

Ask a question! 😊

Submit

### Data Overview (Prompt)

I am researching the relationship between income and sociodemographic census info. I have imported Pandas as `pd`, Numpy as `np`, `scipy`, and `sklearn`, and have a dataframe called `df` loaded into Python. `df` contains the following variables and variable types:

Column Name: age  
Column Type: int64

Column Name: workclass  
Column Type: object  
Sample Values: ['Federal-gov', 'Private', 'Self-emp-inc', 'Local-gov', '?', 'Never-worked', 'State-gov', 'Self-emp-not-inc', 'Without-pay']

Column Name: fnlwgt  
Column Type: int64

Column Name: education  
Column Type: object  
Sample Values: ['10th', '9th', '5th-6th', 'Doctorate', 'Masters', 'Prof-school', 'HS-grad', 'Preschool', 'Some-college', 'Bachelors']

Column Name: education.num  
Column Type: int64

Column Name: marital.status  
Column Type: object  
Sample Values: ['Widowed', 'Divorced', 'Married-civ-spouse', 'Married-spouse-absent', 'Never-married', 'Separated', 'Married-AF-spouse']

Column Name: occupation  
Column Type: object  
Sample Values: ['Tech-support', 'Handlers-cleaners', 'Sales', 'Protective-serv', 'Armed-Forces', 'Exec-managerial', 'Craft-

## Analysis Output



## Data Prompt

```
{DATA_SETUP_INTRO}
```

```
I have imported Pandas as `pd`,  
Numpy as `np`, `scipy`, and  
`sklearn`, and have a dataframe  
called `df` loaded into Python.  
`df` contains the following  
variables and variable types:\n\n
```

```
generateOverview(df)
```

```
{QUESTION}
```

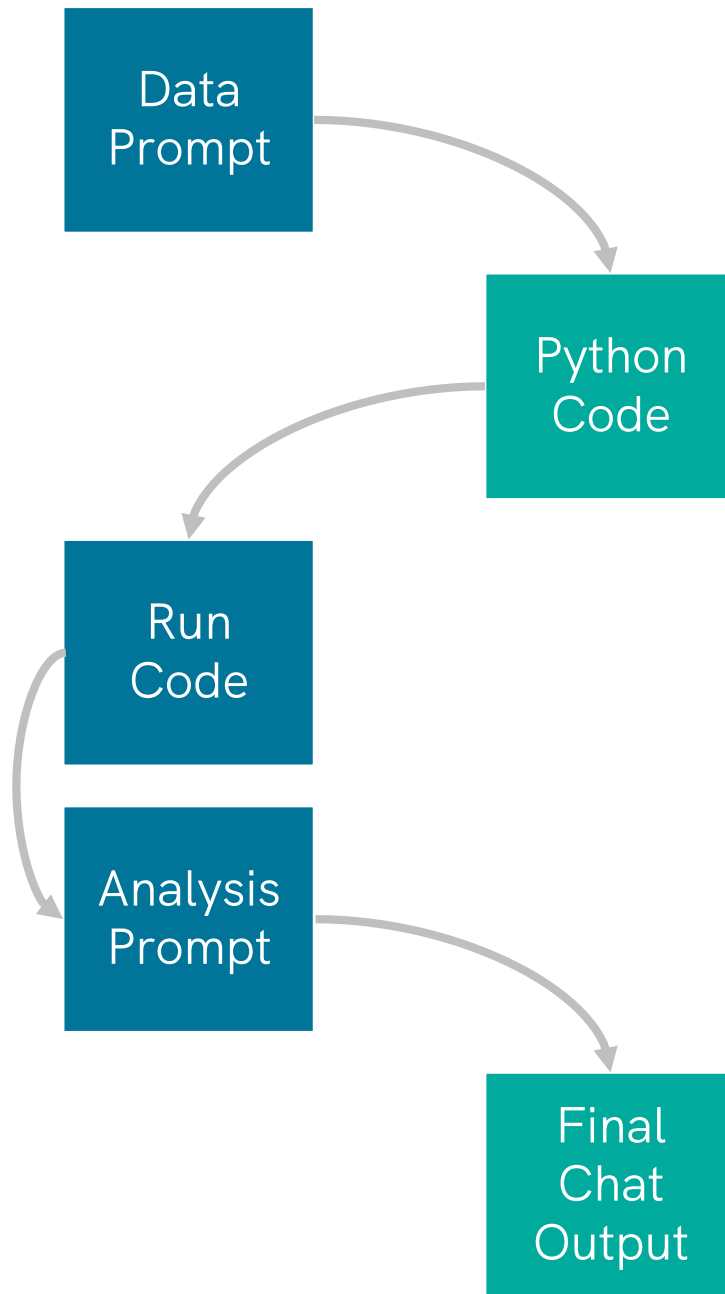
```
I am researching the relationship  
between income and  
sociodemographic census info.
```

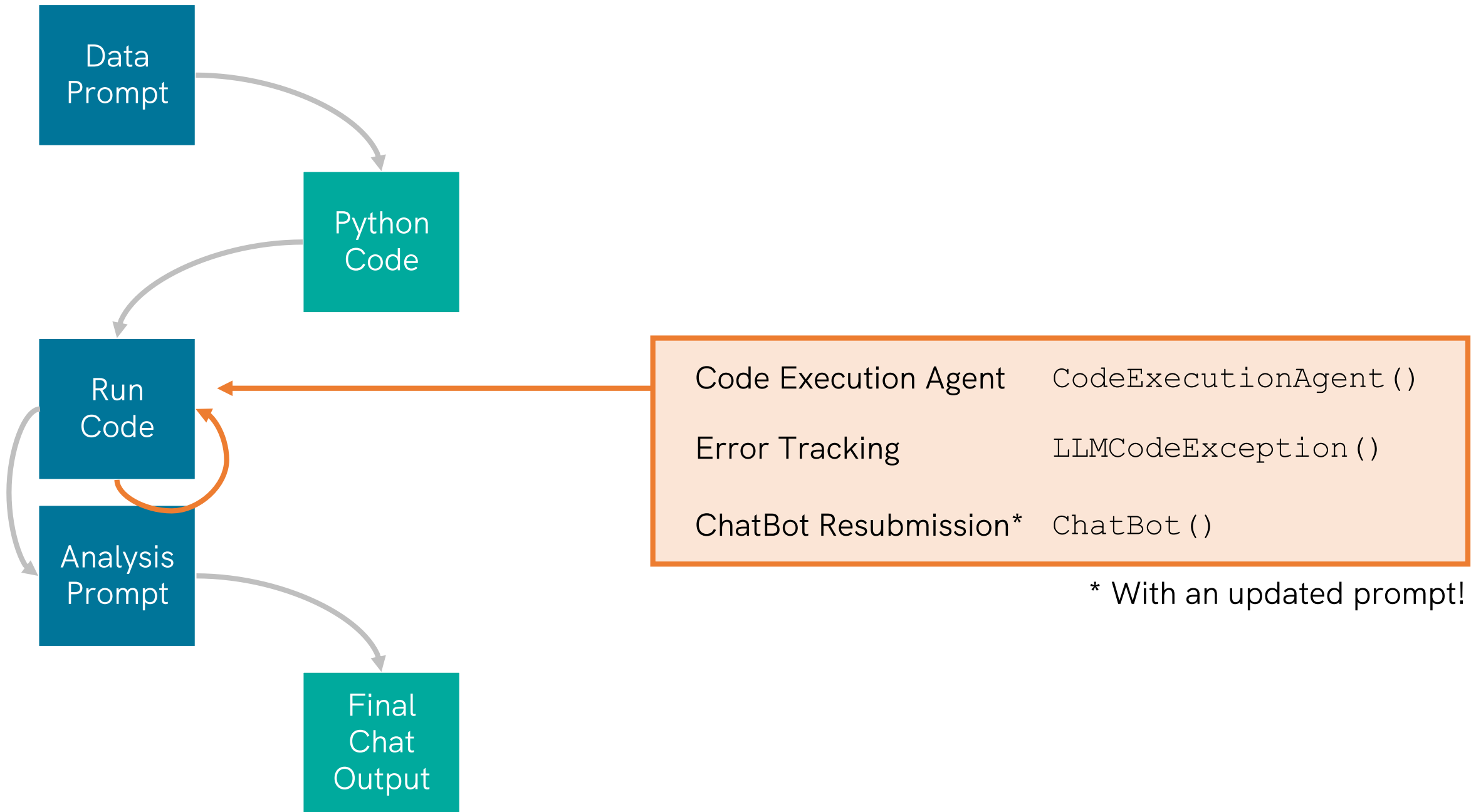
```
I have imported Pandas as `pd`,  
Numpy as `np`, `scipy`, and  
`sklearn`, and have a dataframe  
called `df` loaded into Python.  
`df` contains the following  
variables and variable types:\n\n
```

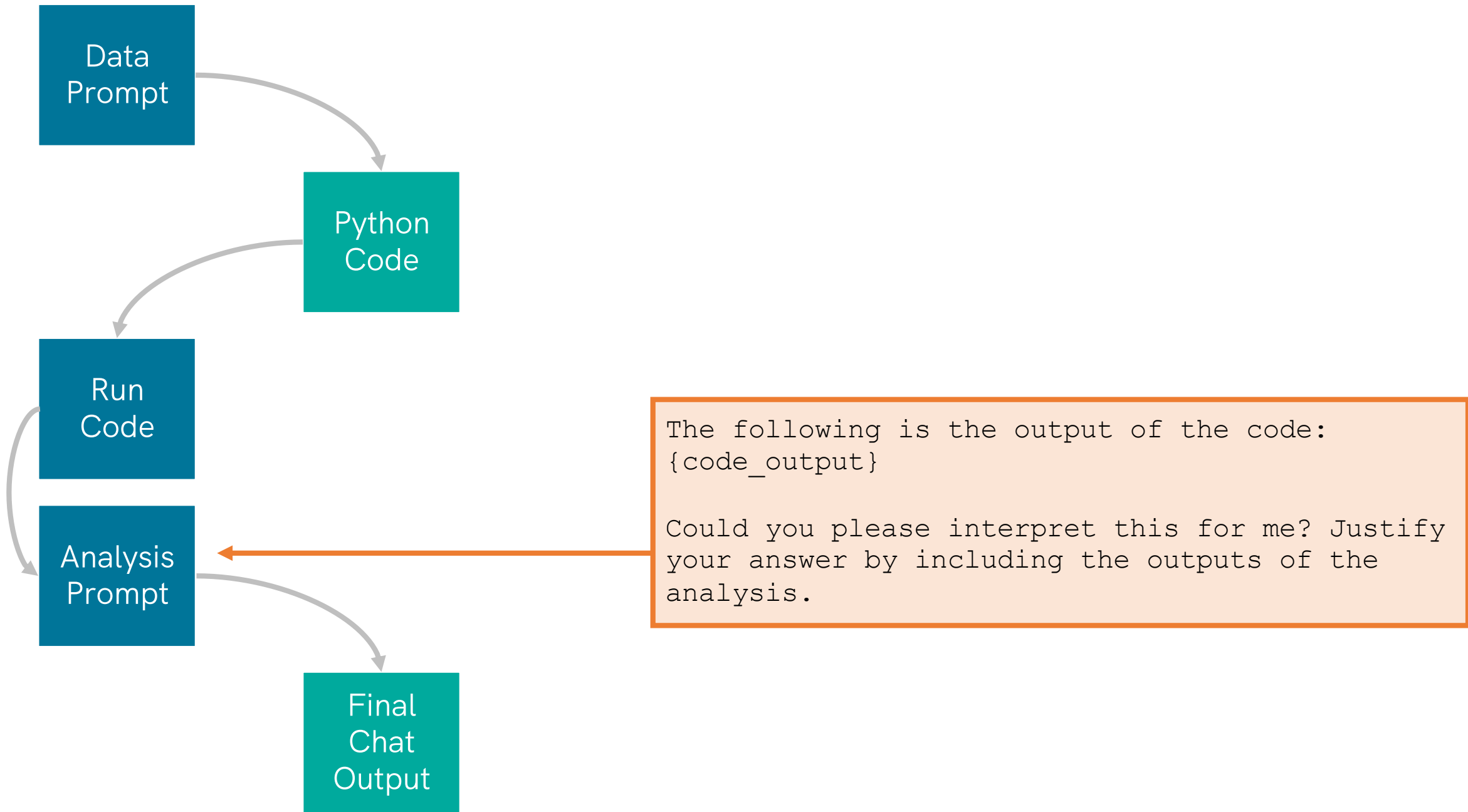
```
Column Name: age  
Column Type: int64
```

```
...
```

```
What is the correlation between  
age and income?
```







## ResearchLLM

Ask a question! 😊

Submit

### Data Overview (Prompt)

I am researching the relationship between income and sociodemographic census info. I have imported Pandas as `pd`, Numpy as `np`, `scipy`, and `sklearn`, and have a dataframe called `df` loaded into Python. `df` contains the following variables and variable types:

Column Name: age  
Column Type: int64

Column Name: workclass  
Column Type: object  
Sample Values: ['Federal-gov', 'Private', 'Self-emp-inc', 'Local-gov', '?', 'Never-worked', 'State-gov', 'Self-emp-not-inc', 'Without-pay']

Column Name: fhlwgt  
Column Type: int64

Column Name: education  
Column Type: object  
Sample Values: ['10th', '9th', '5th-6th', 'Doctorate', 'Masters', 'Prof-school', 'HS-grad', 'Preschool', 'Some-college', 'Bachelors']

Column Name: education.num  
Column Type: int64

Column Name: marital.status  
Column Type: object  
Sample Values: ['Widowed', 'Divorced', 'Married-civ-spouse', 'Married-spouse-absent', 'Never-married', 'Separated', 'Married-AF-spouse']

Column Name: occupation  
Column Type: object  
Sample Values: ['Tech-support', 'Handlers-cleaners', 'Sales',

## Analysis Output

I bet married people work less!! :D

Based on the outputs:

income

<=50K 41.542449

>50K 45.298694

This shows the mean hours worked per week for married individuals in the low (<=50K) vs high (>50K) income groups. On average, married individuals in the high income group work 45.3 hours per week compared to 41.5 hours per week for those in the low income group.

Difference in mean hours worked between high and low earners:

3.756244560088497

The difference between the mean hours worked for the two income groups is 3.8 hours per week.

p-value for difference in mean hours worked: 5.460877332402035e-87

The p-value is very small ( $p < 0.05$ ), indicating the difference in mean hours worked between the two income groups is statistically significant.

The difference in mean hours worked between high and low income married earners is statistically significant ( $p < 0.05$ )

We can conclude that married high earners work significantly more hours per week than married low earners, on average.

In summary, the analysis shows that your hypothesis is correct - married high earners work more hours per week on average compared to married low earners, and this difference is statistically significant. Married high earners work an average of 3.8 more hours per week than married low earners.

Does this help interpret the results? Let me know if you need any clarification

## Next Steps

- Chain of Thought (CoT) reasoning
- Other Agents
- Memory
- Research automation and new knowledge creation

Reach out: [w@phaseai.com](mailto:w@phaseai.com)