FREE Online
Citizen Data Scientist Course

You Too Can be a Citizen Data Scientist – No Matter Your Role, Skill or Job Function

Worksheet

This **Citizen Data Scientist Worksheet** reinforces the ideas, concepts, methods and analytics presented in the course material for the course entitled, 'Be a Citizen Data Scientist – Aspire, Inspire, Become!'. Once you have completed this brief worksheet, you may check your answers using the '**Citizen Data Scientist Worksheet Answer Key**' provided in the supporting documentation.

1. **In your own words, describe the ideal profile of a Citizen Data Scientist candidate**

----

2. **There are numerous benefits inherent in a Citizen Data Scientist initiative. Choose all that apply.**

   Business Users

   Make Timely and Accurate Decisions

   Gain Insight and Perspective

   Share and Collaborate With Others Using Data

   Work on Use Cases Defined by IT or Data Scientists

   Support Day-to-Day Business Decisions

   Produce Analytical Models for Analysis

   Data Scientists

   Reduce Day-to-Day Requests

   Avoid Interaction with Business Users

   Focus on Strategic Projects

   Develop Validated Models for Ongoing Production Use

   Business Enterprise

   Transfer All Data Analytics to IT Staff

   Engender Collaborative Culture

   Reduce Dependence on Data Scientists

   Data and Analytical Democratization

   Reduce Re-Work and Missteps

   Optimize Business Knowledge

3. **When a Citizen Data Scientist initiative is established in a business enterprise, which team members will develop predictive models based on the use cases and utilize domain knowledge. Select one.**

   Data Scientist

   Citizen Data Scientist
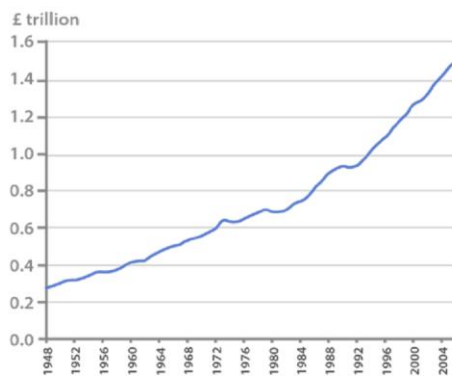
   IT Staff

   Middle Managers

   Executive Team

4. **When a Citizen Data Scientist initiative is established in a business enterprise, which team members will be responsible for strategic analytics when 100% accuracy is required? Choose one!**

   Data Scientist

   Citizen Data Scientist

   IT Staff

   Middle Managers

   Executive Team

5. **Which of the following analytical techniques can be used to analyze Trends and Patterns. Choose up to three.**

   Cyclical

   Clustering

   Seasonal

   Random

   Hierarchical

   Correlation

6. **In this analysis, the Citizen Data Scientist wishes to forecast GDP(x) based on two factors population (a), Index of Industrial Production (IIP) (b), and rainfall (c). Which analytical technique would apply to this type of forecasting?**

Actual GDP (Trillion)	
Years	GDP
Y1	0.35
Y2	0.38
Y3	0.39
Y4	0.40
Y5	0.44
Y6	0.50
Y7	0.58
Y8	0.60
Y9	0.64
Y10	0.70
Forecasted GDP (Trillion)	
Y11	0.82
Y12	0.94
Y13	1.00
Y14	1.22
Y15	1.42

   ARIMAX

   ARIMA

7. **Which analytical technique would a Citizen Data Scientist use to determine the credit limit of a customer based on age, gender, family composition, income and debt ratio.**

    Regression

    Classification

    Clustering

    Correlation

    Hypothesis Testing

8. **Which type of Sampling Analysis would the Citizen Data Scientist use to analyze members in subgroups with similar attributes and characteristics (income, geography, gender, etc.) to analyze demographics.**

    Simple Random Sampling

    Stratified Random Sampling

9. **Under certain circumstances, Outlier analysis will produce results that warrant further exploration. Select the answer that best describes these circumstances.**

    When results illustrate a unique and singular category or niche

    When data points represent items that are far outside the norm

10. **What technique might one use to hypothesize whether the gender of a customer influences the purchase of a certain type of product.**

    Chi Square Test of Independence

    Spearman Rank Correlation

    One Way Anova Test

    Independent Sample T-Test

    Paired Sample T-Test

11. What type of Trend analysis does this illustration depict?

Cyclical

Random

Seasonal

12. Which of these Hypothesis Tests would you use to determine whether there are statistically significant differences between different survey or experiment results.

Chi Square Test of Independence

One Way Anova Test

Independent Sample T-Test

Paired Sample T-Test

13. Which of these methods is used in Descriptive Statistics Analytics. Select all that apply.

Mean

Average

Sensitivity

Median

Gap Analysis

Mode

Range

14. For each of the Business Case Categories listed below, select at least one algorithm technique that would be appropriate to use for analysis.

- **Customer Churn**

Identify loyal customers versus those that will abandon products or services, and rate of retention

Target: Customer Churn – Yes or No

Influencing Factors: Services that each customer uses, customer account information, demographic information about customers

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Product/Service Cross-Selling and Upselling**

Identify product products that can be bundled or combined to increase sales

Influencing Factors: Transaction ID, products purchased together in one transaction

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Medical Costs**

Identify medical cost based on demography of a patient

Target: Medical Cost

Influencing Factors: BMI, gender, age, pre-existing conditions, income, education

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Human Resource Attrition**

  Find the team members and type of employee that will stay/leave based on positions, compensation and other factors

  Target: Employee Attrition – Yes or No

  Influencing Factors: Job satisfaction, satisfaction with pay, performance-reward contingencies, past working experience, employee personal and demographic information

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Marketing Campaign Optimization**

  Identify the most successful email marketing campaigns, the type of customer that opens the email and how often these campaigns results in conversion to sales

  Target: Email Campaign Response – Yes or No

  Influencing Factors: Customer, demography, geography, number of days since last email

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Quality Control**

  Predict quality of output based on process test parameters taken during various stages of manufacturing

  Target: Quality - Low or High

  Influencing Factors: Process testing parameters using measurements from various stages of manufacturing process

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Predictive Analytics Using External Data**

  Determine If the weather and other economic parameters will affect vacancy rates in hotels

  Target: Hotel Vacancies

  Influencing Factors: rainfall, weather parameters, unemployment rate, projected passenger movement into the location

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Customer Segmentation**

  Segment customers to refine target groups for campaigns and promotions

  Influencing Factors: Customer ID, geography, Sales history, discounts, profitability, payment history

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Predict Sales Based on Macro Economic Data**

  Identify the demographic economic factors that affect and sustain sales and predict the sales for various scenarios

  Target: Sales

  Influencing Factors: Rainfall, GDP, Consumer inflation, Wholesale inflation, population growth, last year GDP, last year growth, last year related product (e.g.,sales of cement linked with building material supply) sales growth, this year related product growth projection

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

Citizen Data Scientist Worksheet Answer Key

PLEASE COMPLETE THE WORKSHEET

THEN

USE THIS ANSWER KEY TO VERIFY YOUR ANSWERS

1. **In your own words, describe the ideal profile of a Citizen Data Scientist candidate**

   Gartner defines a *Citizen Data Scientist* as 'a person who creates or generates models that leverage predictive or prescriptive analytics but whose primary job function is outside of the field of statistics and analytics.' A Citizen Data Scientist is different from a true Data Scientist in one crucial way; namely, they do not have the skills or training to be an analyst or a programmer but, with the right tools, they are capable of generating reports, analyzing data and sharing data to make decisions

2. **There are numerous benefits inherent in a Citizen Data Scientist initiative. Choose all that apply.**

   Business Users

   Make Timely and Accurate Decisions   Work on Use Cases Defined by IT or Data Scientists

   Gain Insight and Perspective   Support Day-to-Day Business Decisions

   Share and Collaborate With Others Using Data   Produce Analytical Models for Analysis

   Data Scientists

   Reduce Day-to-Day Requests   Focus on Strategic Projects

   Avoid Interaction with Business Users   Develop Validated Models for Ongoing Production Use

   Business Enterprise

   Transfer All Data Analytics to IT Staff   Data and Analytical Democratization

   Engender Collaborative Culture   Reduce Re-Work and Missteps

   Reduce Dependence on Data Scientists   Optimize Business Knowledge

3. **When a Citizen Data Scientist initiative is established in a business enterprise, which team members will develop predictive models based on the use cases and utilize domain knowledge. Select one.**

   Data Scientist

   Citizen Data Scientist

   IT Staff

   Middle Managers

   Executive Team

4. **When a Citizen Data Scientist initiative is established in a business enterprise, which team members will be responsible for strategic analytics when 100% accuracy is required? Choose one!**

    Data Scientist

    Citizen Data Scientist

    IT Staff

    Middle Managers

    Executive Team

5. **Which of the following analytical techniques can be used to analyze Trends and Patterns. Choose up to three.**
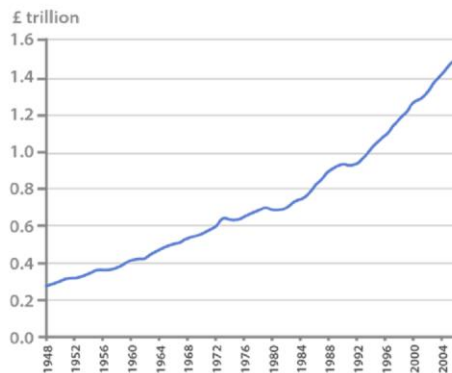
    Cyclical

    Clustering

    Seasonal

    Random

    Hierarchical

    Correlation

6. **In this analysis, the Citizen Data Scientist wishes to forecast GDP(x) based on two factors population (a), Index of Industrial Production (IIP) (b), and rainfall (c). Which analytical technique would apply to this type of forecasting?**

Actual GDP (Trillion)	
Years	GDP
Y1	0.35
Y2	0.38
Y3	0.39
Y4	0.40
Y5	0.44
Y6	0.50
Y7	0.58
Y8	0.60
Y9	0.64
Y10	0.70
Forecasted GDP (Trillion)	
Y11	0.82
Y12	0.94
Y13	1.00
Y14	1.22
Y15	1.42

ARIMAX

ARIMA

7. **Which analytical technique would a Citizen Data Scientist use to determine the credit limit of a customer based on age, gender, family composition, income and debt ratio.**

    <mark>Regression</mark>

    Classification

    Clustering

    Correlation

    Hypothesis Testing

8. **Which type of Sampling Analysis would the Citizen Data Scientist use to analyze members in subgroups with similar attributes and characteristics (income, geography, gender, etc.) to analyze demographics.**

    Simple Random Sampling

    <mark>Stratified Random Sampling</mark>

9. **Under certain circumstances, Outlier analysis will produce results that warrant further exploration. Select the answer that best describes these circumstances.**

    When results illustrate a unique and singular category or niche

    <mark>When data points represent items that are far outside the norm</mark>

10. **What technique might one use to hypothesize whether the gender of a customer influences the purchase of a certain type of product.**

    <mark>Chi Square Test of Independence</mark>

    Spearman Rank Correlation

    One Way Anova Test

    Independent Sample T-Test

    Paired Sample T-Test

11. **What type of Trend analysis does this illustration depict?**

<mark>Cyclical</mark>

Random

Seasonal

12. **Which of these Hypothesis Tests would you use to determine whether there are statistically significant differences between different survey or experiment results.**

Chi Square Test of Independence

One Way Anova Test

Independent Sample T-Test

<mark>Paired Sample T-Test</mark>

13. **Which of these methods is used in Descriptive Statistics Analytics. Select all that apply.**

<mark>Mean</mark>

Average

Sensitivity

<mark>Median</mark>

Gap Analysis

<mark>Mode</mark>

<mark>Range</mark>

14. **For each of the Business Case Categories listed below, select at least one algorithm technique that would be appropriate to use for analysis.**

- **Customer Churn**

Identify loyal customers versus those that will abandon products or services, and rate of retention

Target: Customer Churn – Yes or No

Influencing Factors: Services that each customer uses, customer account information, demographic information about customers

Trends and Patterns	Sampling	Outliers	Forecasting
<mark>Classification</mark>	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Product/Service Cross-Selling and Upselling**

Identify product products that can be bundled or combined to increase sales

Influencing Factors: Transaction ID, products purchased together in one transaction

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
<mark>Frequent Pattern Mining</mark>	Hypothesis Testing	Descriptive Statistics	

- **Medical Costs**

Identify medical cost based on demography of a patient

Target: Medical Cost

Influencing Factors: BMI, gender, age, pre-existing conditions, income, education

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	<mark>Regression</mark>
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Human Resource Attrition**

  Find the team members and type of employee that will stay/leave based on positions, compensation and other factors

  Target: Employee Attrition – Yes or No

  Influencing Factors: Job satisfaction, satisfaction with pay, performance-reward contingencies, past working experience, employee personal and demographic information

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Marketing Campaign Optimization**

  Identify the most successful email marketing campaigns, the type of customer that opens the email and how often these campaigns results in conversion to sales

  Target: Email Campaign Response – Yes or No

  Influencing Factors: Customer, demography, geography, number of days since last email

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Quality Control**

  Predict quality of output based on process test parameters taken during various stages of manufacturing

  Target: Quality - Low or High

  Influencing Factors: Process testing parameters using measurements from various stages of manufacturing process

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Predictive Analytics Using External Data**

  Determine If the weather and other economic parameters will affect vacancy rates in hotels

  Target: Hotel Vacancies

  Influencing Factors: rainfall, weather parameters, unemployment rate, projected passenger movement into the location

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	==Regression==
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Customer Segmentation**

  Segment customers to refine target groups for campaigns and promotions

  Influencing Factors: Customer ID, geography, Sales history, discounts, profitability, payment history

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	==Clustering==	Correlation	Regression
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	

- **Predict Sales Based on Macro Economic Data**

  Identify the demographic economic factors that affect and sustain sales and predict the sales for various scenarios

  Target: Sales

  Influencing Factors: Rainfall, GDP, Consumer inflation, Wholesale inflation, population growth, last year GDP, last year growth, last year related product (e.g.,sales of cement linked with building material supply) sales growth, this year related product growth projection

Trends and Patterns	Sampling	Outliers	Forecasting
Classification	Clustering	Correlation	==Regression==
Frequent Pattern Mining	Hypothesis Testing	Descriptive Statistics	