Pleaides Whitepaper

An Interactive Exploratory Data Analysis Framework

By: Dominic Zygadlo

Initial Proposal

Use Case: A person (for the intents of storytelling we shall name her Lux) wants to analyze this interesting dataset that she found; however she is stuck because of executive dysfunction. She has experience analyzing data and figures, however getting from importing the data to a model that she can derive results from is daunting. Here comes in Pleiades. Lux imports her csv file into the dashboard, and she can interactively choose what kind of analysis she wants to pursue. Even if she does not know where to being, Pleiades provides a general roadmap for different types of analyses, as well as describing how to explain the more “black-box” tools.

1. Lux the data into the Dashboard
2. Immediately she can view the data in a Table View
    a. Each column displays the Name and data type
    b. Column menus, clicking on the downward button, Lux may view the list of operations available against this column
        1. Rename
        2. Filter
            a. Equal to …
            b. Not equal to …
            c. Is in …
            d. Is not in …
            e. Less than …
            f. Less than or equal to …
            g. Greater than …
            h. Greater than or equal to …
            i. Between …
            j. Not between …
            k. Not NA …
            l. Not Outliers …
            m. Is Outliers …
        3. Create Bins (Categories)
        4. Create Calculation (Mutate)
        5. Arrange (Sort)
        6. Change data type
3. By clicking on the Summary tab, Lux may view basic descriptive statistics for each variable in an interactable (each cell is moveable) grid format
    a. Distribution chart with button to expand view
    b. Metrics
        i. Minimum
        ii. Maximum
        iii. Median
        iv. Mean
        v. Standard Deviation
        vi. Quartile 1, Quartile 3
        vii. NA Count
        viii. Unique Count
    c. Highlight Mode allows Lux to select specific constraints on the data to compare underlying distributions and metrics with the sample level. There is an exit button
    d. Correlation Mode allows Lux to view the correlations of a target column against the other data
        i. Pearson
        ii. Spearman
    e. Change Sort Order

i. R square
ii. Absolute Correlation
iii. P value
iv. AUC

4. Finally on the Analytics Tab, Lux can create the specific charts for a variety of statistical tests. She can also begin fitting the data to basic machine learning algorithms
   a. Statistical Tests
      i. T test (only if a comparing column has 2 unique values)
      ii. ANOVA (only if a comparing column has 2 unique values)
      iii. Wilcoxon Test
      iv. Kruskal-Wallis Test
      v. Chi-Square Test
   b. Machine Learning Algorithms
      i. Dimensionality Reduction
         1. Principal Component Analysis
      ii. Regression
         1. Simple Linear
         2. Multiple Linear
         3. Generalized Linear
         4. LASSO
         5. Ridge Regression
         6. LARS Lasso
         7. Stochastic Gradient Descent
         8. Support Vector Machine
      iii. Classification
         1. Logistic Regression
         2. Naïve Bayes
         3. Decision Tree
         4. Random Forest
         5. XGBoost
         6. Support Vector Machine
      iv. Clustering
         1. K-Means
         2. Spectr4al clustering
         3. DBSCAN
         4. OPTICS
         5.

Data Import

File Data

You can import both local and remote data quickly.

1. Click the **+** button next to **Data Frames** and select **Import File Data**
2. If you are importing files from your local device (e.g. desktop, laptop, etc.), then make sure the **Local** tab is selected
    a. Click on the corresponding file type icon from the following options: **Text File (.csv)**, **Excel File (.xlsx), JSON File (.json), Pickle File (.pickle)**
    b. If you select local, a file picker window will pop-up to select a file for importing
3. If you are importing files from a remote device (e.g. URL), then make sure the **Remote** tab is selected
    a. Type in or paste the remote file's URL in the **first text box**
    b. Select the file type from the dropdown of options: **Text File (.csv)**, **Excel File (.xlsx), JSON File (.json), Pickle File (.pickle)**
4. When importing files, users may select different **input parameters** to customize how the data frame will be built
    a. **Separator** – character(s) used to separate fields within a record
    b. **Treat as NA** – strings or values that should be treated as missing values, a **comma** should be used to enter multiple strings
    c. **First Row as Header** – True or False, if False then column names will be alphabetic
    d. **Column Types** – users may override column types and names by directly click on the column in the displayed data frame head, column types may be adjusted by entering **col_type = dtype**
        i. **N** – nominal data (e.g. color, gender, ethnicity, etc.)
        ii. **O** – ordinal data (e.g. placing, letter grading, economic status, etc.)
        iii. **D** – discrete data (e.g. counts)
        iv. **C** – continuous data (e.g. price, time, measurements, etc.)
    e. **Column Selection** – users may rename and/or include/exclude columns by clicking on the **column name** and/or **checkbox** next to the desired column

Database Data

You can create connections to existing databases to query data. Examples include, **Google BigQuery**, **Snowflake**, **MySQL DB**, **PostgreSQL**

WebApp Data

You can scrape data directly from **Twitter**, **GitHub**, or even **Web Pages**, with our in-house data mining algorithms.

Data Wrangling

Command Line Mode

Users can enable command line mode if they prefer directly typing in their own **pandas** and **NumPy** functions.

Feature Manipulation

Users will have access to a variety of methods to clean and customize their dataset.

1. **Select / Remove Columns**
2. **Reorder Columns / Rows**
3. **Create New Calculation(s)**
4. **Filter**

5. **Rename**
6. **Join / Bind / Union / Intersection / Difference**
7. **Unique Only**
8. **Drop NA**
9. **Train / Test / Validation Split**
10. **One hot encoding**

Visualization

Numeric

1. ONE Numeric
    a. Histogram
    b. Density Plot
2. TWO Numeric
    a. Not Ordered
        i. Box Plot
        ii. Violin Plot
        iii. Histogram
        iv. Density Plot
        v. Scatter Plot
        vi. 2D Density Plot
    b. Ordered
        i. Connected Scatter Plot
        ii. Area Plot
        iii. Line Plot
3. THREE Numeric
    a. Not Ordered
        i. Box Plot
        ii. Violin Plot
        iii. Bubble Plot
        iv. 3D Scatter or Surface
    b. Ordered
        i. Stacked Area Plot
        ii. Stream Graph
        iii. Line Plot
        iv. Area (SM)
4. FOUR+ Numeric
    a. Not Ordered
        i. Box Plot
        ii. Violin Plot
        iii. Ridge Line
        iv. PCA
        v. Correlogram
        vi. Heatmap
        vii. Dendrogram
    b. Ordered
        i. Stacked Area Plot
        ii. Stream Graph
        iii. Line Plot
        iv. Area (SM)

Categorical

1. ONE Categorical
    a. Bar Plot
    b. Lollipop
    c. Waffle
    d. Word Cloud
    e. Doughnut
    f. Pie
    g. Tree Map
    h. Circular Packing
2. TWO+ Categorical
    a. Independent Lists
        i. Venn Diagram
    b. Nested
        i. Tree Map
        ii. Circular Packing
        iii. Sunburst
        iv. Bar Plot
        v. Dendrogram
    c. Subgroup
        i. Grouped Scatter
        ii. Heat Map
        iii. Lollipop
        iv. Grouped Bar Plot
        v. Stacked Bar Plot
        vi. Parallel Plot
        vii. Spider Plot
        viii. Sankey Diagram
    d. Adjacency
        i. Network
        ii. Chord
        iii. Arc
        iv. Sankey
        v. Heatmap

Multivariate

1. ONE Numeric + ONE Categorical
    a. One observation per group
        i. Boxplot
        ii. Lollipop
        iii. Doughnut
        iv. Pie
        v. Word Cloud
        vi. Tree Map
        vii. Circular Packing
        viii. Waffle
    b. Several observations per group
        i. Box Plot
        ii. Violin
        iii. Ridge Line
        iv. Density
        v. Histogram

2. TWO+ Numeric + ONE Categorical
   a. No Order
      i. Grouped Scatter
      ii. 2D Density
      iii. Box Plot
      iv. Violin
      v. PCA
      vi. Correlogram
   b. Ordered Number
      i. Stacked Area
      ii. Area
      iii. Steam Graph
      iv. Line Plot
      v. Connected Scatter
   c. One Value per Group
      i. Grouped Scatter
      ii. Heat Map
      iii. Lollipop
      iv. Grouped Bar Plot
      v. Stack Bar Plot
      vi. Parallel Plot
      vii. Spider Plot
      viii. Sankey Diagram
3. One Numeric + TWO+ Categorical
   a. Subgroup
      i. One Observation per Group
         1. Grouped Scatter
         2. Heat Map
         3. Lollipop
         4. Grouped Bar Plot
         5. Stack Bar Plot
         6. Parallel Plot
         7. Spider Plot
         8. Sankey Diagram
      ii. Two+ Observations per Group
         1. Box Plot
         2. Violin
   b. Nested
      i. One Observation per Group
         1. Bar Plot
         2. Dendrogram
         3. Sunburst
         4. Tree Map
         5. Circular Packing
      ii. Two+ Observations per Group
         1. Box Plot
         2. Violin
   c. Adjacency
      i. Network
      ii. Chord
      iii. Arc
      iv. Sankey

v. Heatmap

Bar Charts

Inputs:

1. **Orientation** – either vertical or horizontal
2. **X-axis Feature**
3. **Y-axis Feature**
4. **Color by Feature**
5. **Sort by Feature**
6. **Repeat by Feature**
7. **Stack or Group by Feature(s)**
8. **Highlight**
9. **Reference Line(s)**

Use Cases:

1.

Line Charts

Inputs:

1. **X-axis Feature**
2. **Y-axis Feature**
3. **Color by Feature**
4. **Sort by Feature**
5. **Repeat by Feature**
6. **Marker Type**
7. **Highlight**
8. **Range**
9. **Reference Line(s)**
10. **Trendline(s)**

Use Cases:

1.

Area Charts

Inputs:

1. **X-axis Feature**
2. **Y-axis Feature**
3. **Color by Feature**
4. **Sort by Feature**
5. **Repeat by Feature**
6. **Marker Type**
7. **Highlight**

Use Cases:

1.

Pie / Ring Charts

Inputs:

1. **Value Feature**
2. **Sort by Feature**
3. **Repeat by Feature**
4. **Style**
5. **Highlight**

Use Cases:

Histograms

Inputs:

1. **X-ax Feature**
2. **Color by Feature**
3. **Number of Bars**
4. **Repeat by Feature**
5. **Highlight**
6. **Cumulative Sum Reference Line**

Use Cases:

Density Plots

Inputs:

1. **X-axis Feature**
2. **Color by Feature**
3. **Repeat by Feature**
4. **Include Outlier**

Use Cases:

Boxplots

Inputs:

1. **X-axis Feature**
2. **Y-axis Feature**
3. **Color by Feature**
4. **Repeat by Feature**
5. **Sort By Feature**
    a. **Sum**
    b. **Median**
    c. **Min**
    d. **Max**
    e. **IQR**
    f. **Standard Deviation**
6. **Outlier Detection**

Use Cases:

Violin Plots

Inputs:

1. **X-axis Feature**
2. **Y-axis Feature**
3. **Color by Feature**
4. **Repeat by Feature**
5. **Sort By Feature**
    a. **Sum**
    b. **Median**
    c. **Min**
    d. **Max**
    e. **IQR**
    f. **Standard Deviation**
6. **Outlier Detection**
7. **Include Boxplot / Dotplot**

Use Cases:

Analytics

Correlations

Users can display correlations by Feature

1. **Pearson**
2. **Kendall**
3. **Spearman**

Inputs:

1. **Selected Variables**
    a. **X-axis Feature(s)**
    b. **Y-axis Feature(s)**
2. **Color by Feature**
3. **Repeat by Feature**
4. **Positive Only**
5. **Negative Only**

Outputs:

1. **Correlogram in Descending Order**
2. **Scatter Matrix in Descending Order**

Principal Component Analysis

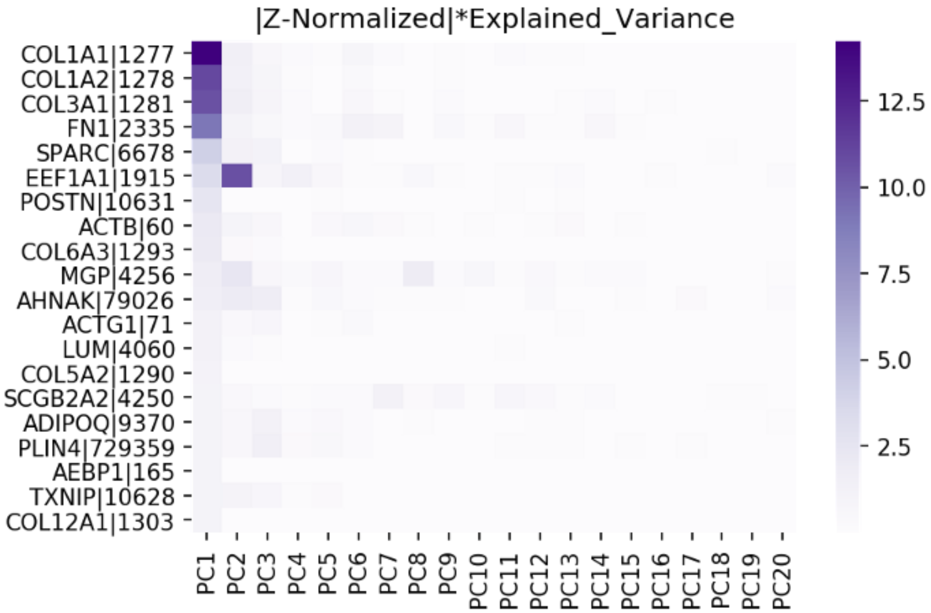
Inputs:

1. **Selected Numeric Features**
2. **Color by Features**
3. **Kernel** – linear; polynomial; radial basis function; sigmoid; cosine; precomputed

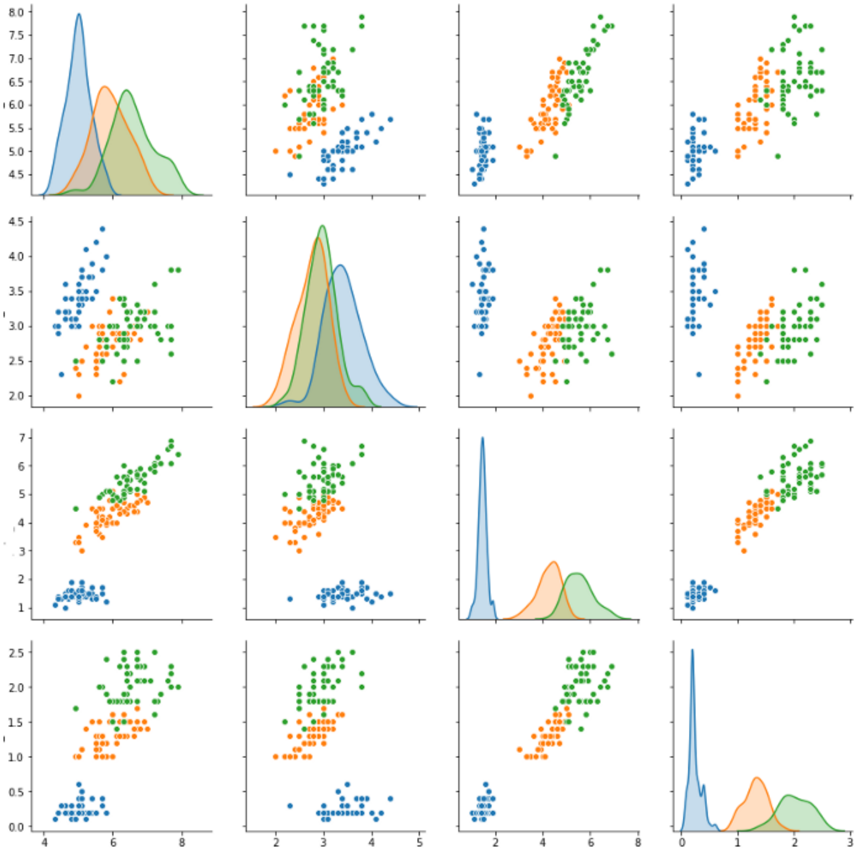
Outputs:

1. **Data frame with Principal Components**
2. **Scoring of Principal Component Features** – displays the top n features based on their absolute eigenvalue contributions to the individual principal components

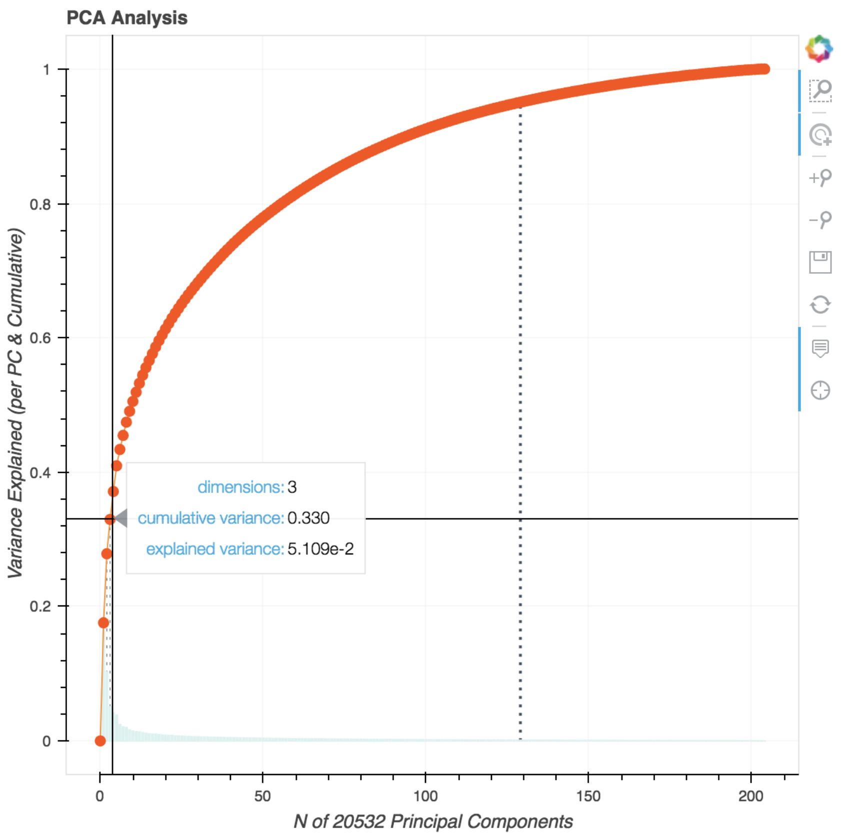
3. **Heatmap of Explained Variance**



4. **Scatter Matrix Sorted by Descending Explained Variance**



5. **Cumulative Explained Variance Plot with 95% Highlight**



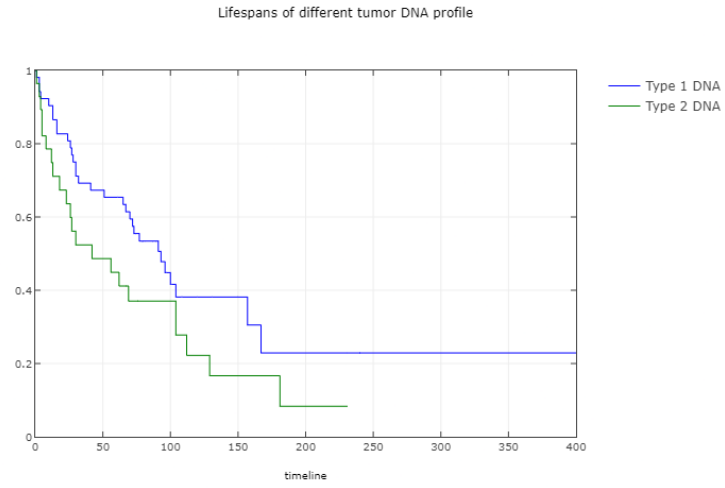
Survival Estimator

Inputs:

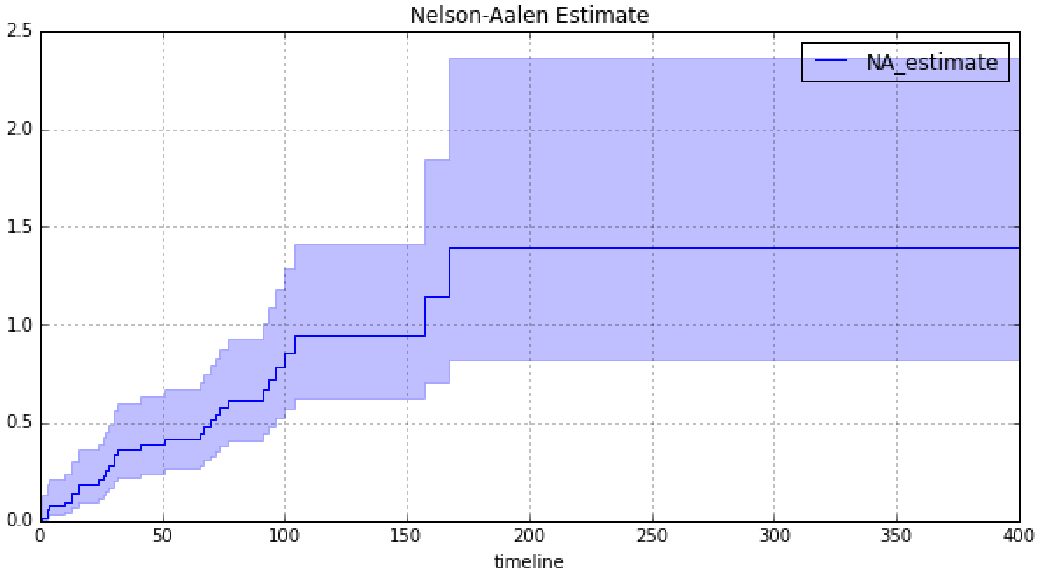
1. **Start time**
2. **End time**
3. **Event Feature**
4. **Color by Feature(s)**
5. **Confidence Interval**

Outputs:

1. Kaplan-Meier Curve



2. Nelson-Aalen Hazard Curve



Hypothesis Testing

1. **T Test**
2. **ANOVA**
3. **Wilcoxon Test**
4. **Kruskal-Wallis Test**
5. **Chi-Square Test**
6. **A/B Test**
7. **Normality Test**
    a. **Anderson-Darling Test for Error Normality**
    b. **Shapiro-Wilk Test for Error Normality**
8. **Variance Inflation Test**
9. **Outlier Detection**
    a. **Normalized Quartile Fences**
    b. **DBSCAN**
10. **Constant Error Variance Test**
    a. **Brown-Forsythe Test**
    b. **Breusch-Pagan Test**