**Microsoft Released Data Science Utilities V0.12 for Team Data Science Process**

Microsoft released [Team Data Science Process (TDSP)](https://github.com/Azure/Microsoft-TDSP) in September 2016 along with a set of data science [utilities](https://blogs.technet.microsoft.com/machinelearning/2016/10/18/two-new-utilities-to-boost-your-data-science-productivity/) (V0.1) aimed to boost data science productivity. Now [V0.12 of TDSP utilities](https://github.com/Azure/Azure-TDSP-Utilities) is released, with the following new features and enhanced features.

**New Features:**

* **IDEAR in MRS for Big Data**

Microsoft R Server (MRS) is the enterprise-class analytics platform for R. It supports exploring, visualizing, and analyzing big data on single machine, Hadoop clusters, or Spark clusters. The previously released IDEAR in open source R is constrained by the memory size as the data is loaded into the memory before data exploration. Now we released IDEAR in MRS to allow R users to explore and analyze big data interactively, and to generate data reports automatically. The feature changes are mostly under the hood, not quite visible from the UI. Therefore, IDEAR in MRS brings the same user experience as IDEAR in open source R with extended capability in handling big data. Microsoft offers free [Microsoft R Server Developer Version](https://blogs.technet.microsoft.com/machinelearning/2016/01/12/making-r-the-enterprise-standard-for-cross-platform-analytics-both-on-premises-and-in-the-cloud/). If you are using an Azure [Data Science Virtual Machine](http://aka.ms/dsvmhandout), MRS developer version is already installed. You can start using IDEAR in MRS instantly.

* **IDEAR in Python 3**

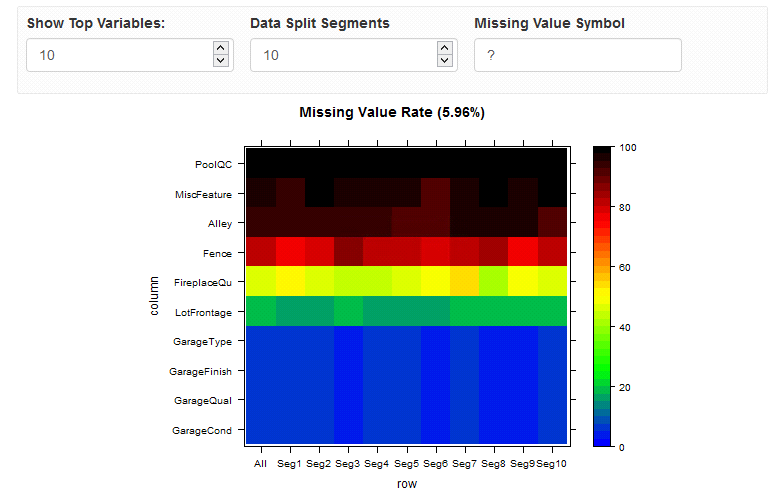
Since [Python 2.7 will not be maintained past 2020](https://pythonclock.org/), it makes sense to develop IDEAR in Python 3. The newly released IDEAR in Python can run in both Python 3.5 and Python 2.7. Our future Python version of IDEAR will be only on Python 3.x. IDEAR in Python 2.7 will be depreciated.

* **IDEAR in Python 3 on Azure Notebooks Services**

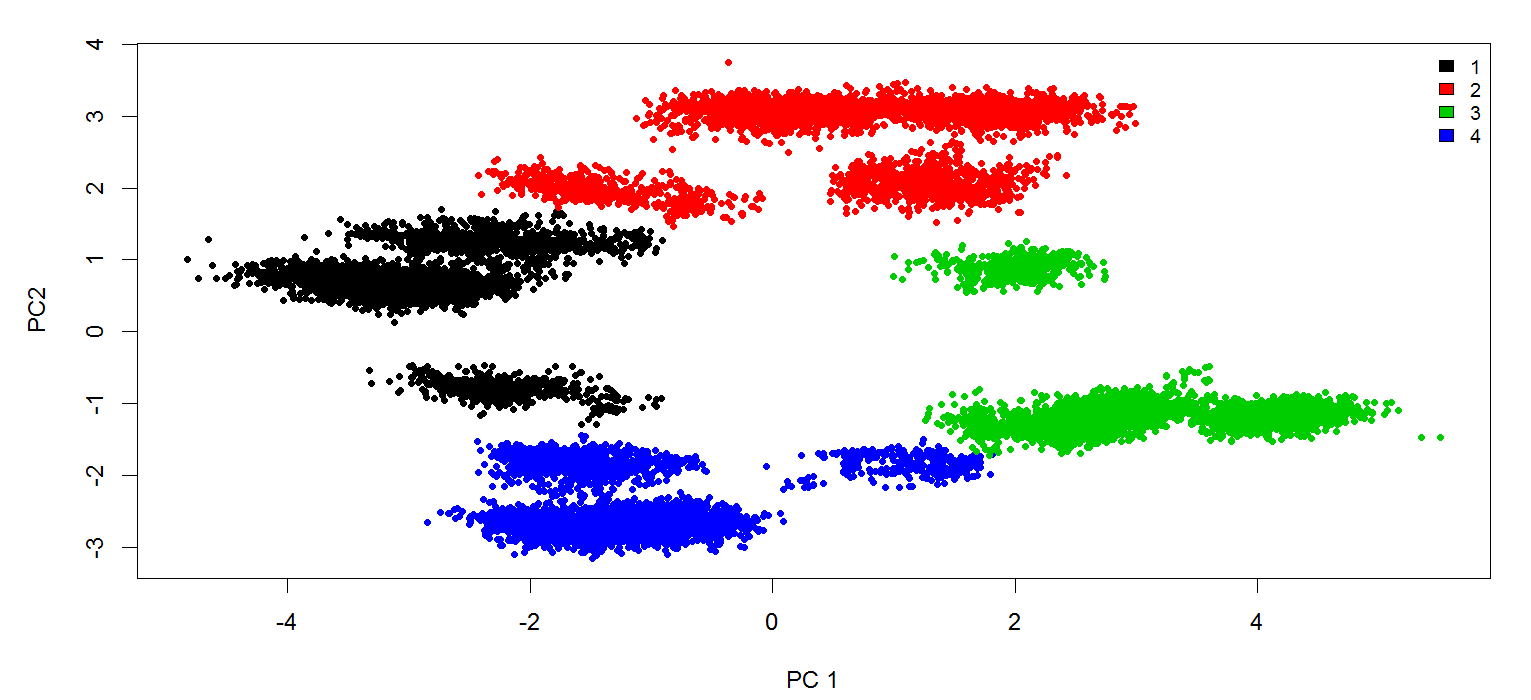
We also released an [Azure Notebooks](https://notebooks.azure.com) service version of IDEAR in Python 3.5, named ***IDEAR-Python-AzureNotebooks.ipynb***. Using the Azure Notebook services can save you a lot of time on setting up Jupyter Notebook servers by yourself and installing necessary libraries. ***IDEAR-Python-AzureNotebooks.ipynb*** reads both data and YAML files from Azure Blob Storage. All interactive data exploration, analysis, and visualization capabilities are the same as the IDEAR in Jupyter Notebooks (***IDEAR.ipynb***). The only difference is that ***IDEAR-Python-AzureNotebooks.ipynb*** does not have functions to generate report automatically.

**Enhanced Features**:

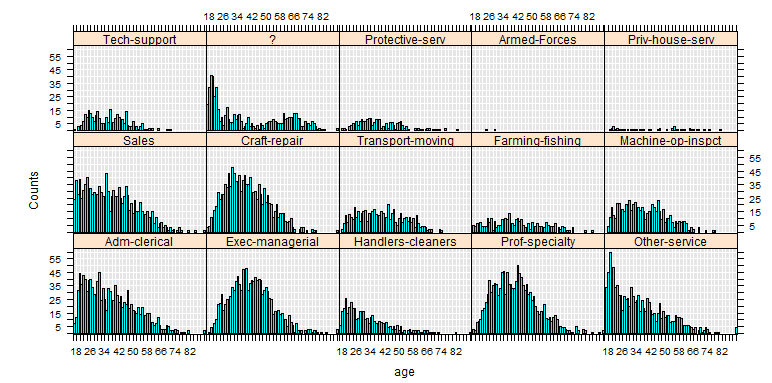
* **Checking missing values in IDEAR in R**. Missing value is one of the data quality issues data scientist pay close attention to when doing data analysis. Now we provide a feature to assess and visualize the severity of missing values in the data. It helps users to identify which variables have the highest rates of missing values, and where missing values happen (which segments of rows).



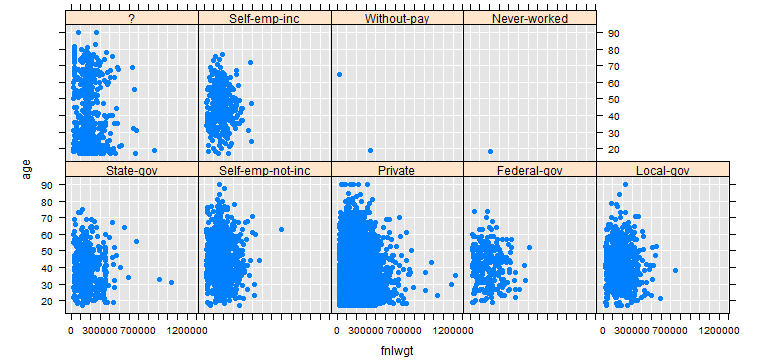
* **Principal component analysis on mixture of data types in IDEAR (open source R)**. It is almost universally true that both numerical and categorical variables co-exist in every data set. Sometime even the categorical variables dominate the data set. In this release, we used [PCAmixdata](https://cran.r-project.org/web/packages/PCAmixdata/PCAmixdata.pdf) to handle mixture of categorical and numerical variables. The following figure demonstrates a clear clustering pattern, colored by variable *season*, when applying IDEAR on the Bike Rental sample data shipped with the utilities by using the PCAmixdata library.



* **Added numerical variable histograms grouped by categorical variable levels in IDEAR in MRS.** The allows users to easily compare the distribution difference of a numerical variable conditioning on different values of the categorical variable.



* **Added numerical interactions grouped by categorical variables in IDEAR in MRS.** Interactions between two numerical variables could be influenced by a third categorical variable. Now you have the option to view the scatterplot between numerical variables grouped by the categorical variable levels.



**Next Steps**

You can download and play with these new features in the data science utilities. Don’t hesitate to send us feedbacks and feature requests via the comments feature below, or on the [**issues tab**](https://github.com/Azure/Azure-TDSP-Utilities/issues) of our GitHub repository, or via tweet to @zenlytix. We are working to improve our tools to better serve your data science project needs.

**References**

<https://msdn.microsoft.com/en-us/microsoft-r>

<https://blogs.technet.microsoft.com/machinelearning/2016/10/11/introducing-the-team-data-science-process-from-microsoft/>

<https://blogs.technet.microsoft.com/machinelearning/2017/01/11/announcing-data-science-utilities-version-0-11-for-the-team-data-science-process/>