Department of the Air Force

Cloud One Application Migration Cost Model – FY22 Update

Vinny Papia, GS-14
AFLCMC/HNF

Sept 2022

Version 6

ENTERPRISE IT + CYBER INFRASTRUCTURE

U.S. Air Force 75 Established 1947

Agenda

- The Problem
- What is Cloud One and where did this data come from?
- Data Collection and Preparation
- Data Cleansing and Variable Preparation
- FY21 Model and Issues Encountered
- FY22 Data Updates and Model Testing and Results
- FY22 Model and Statistics
- Future Work

IT + CYBER

The Problem

- Cloud One is now managing migration services for more than 100 applications; other Services are large-scale cloud migration projects
- Currently, app owners wait until application analysis is completed before receiving an accurate estimate of their migration costs
- Application analysis typically requires a $70k-$100k investment, which is significant for many smaller applications
- **Decision makers need insight into the cost of their application migrations prior to making the decision to migrate**

Cloud One Background

Cloud One is the Air Force's commercial cloud environment and a program office offering application migration services

In 2017, the Air Force Common Computing Environment (CCE), Cloud One's predecessor, attempted a new cloud deployment model:

- Use a prime contractor as a lead only, mainly providing management services
- Create a team of expert cloud migration subcontractors
- Test out the skills of each subcontractor by awarding them a simple application migration
- If successful, award the subcontractor additional migrations, if unsuccessful, cut them from the team

The team involved in building this process recognized the need to collect data to help inform the future of the program, so we did!

While cost data was specifically collected during the initial stages, substantial technical data was available as well, providing a full picture of many application migrations

Initial Model Data Collection and Preparation Overview

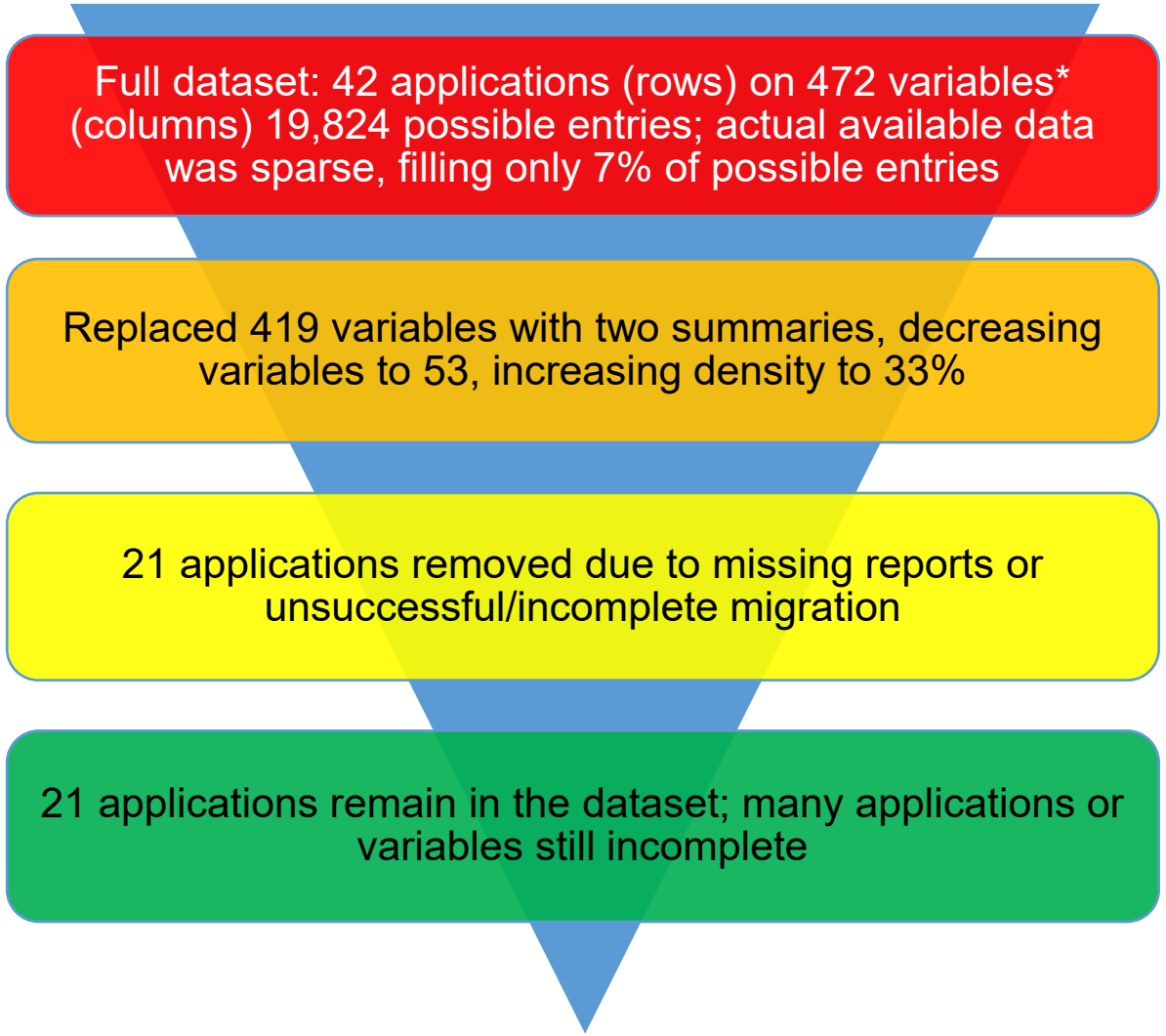
85 application analysis reports collected in PDF and PowerPoint format + migration costs (from CTR financial reports, normalized to CY20$) for 42 applications

Manually extracted application details from reports (lines of code, current OS, users, etc.), 472 variables per application

30 applications removed from model development due to reports missing key details or incomplete migrations

12 applications with 7 variables per application

Data Cleansing and Variable Preparation



*full variable list in backup

Data Cleansing and Variable Preparation

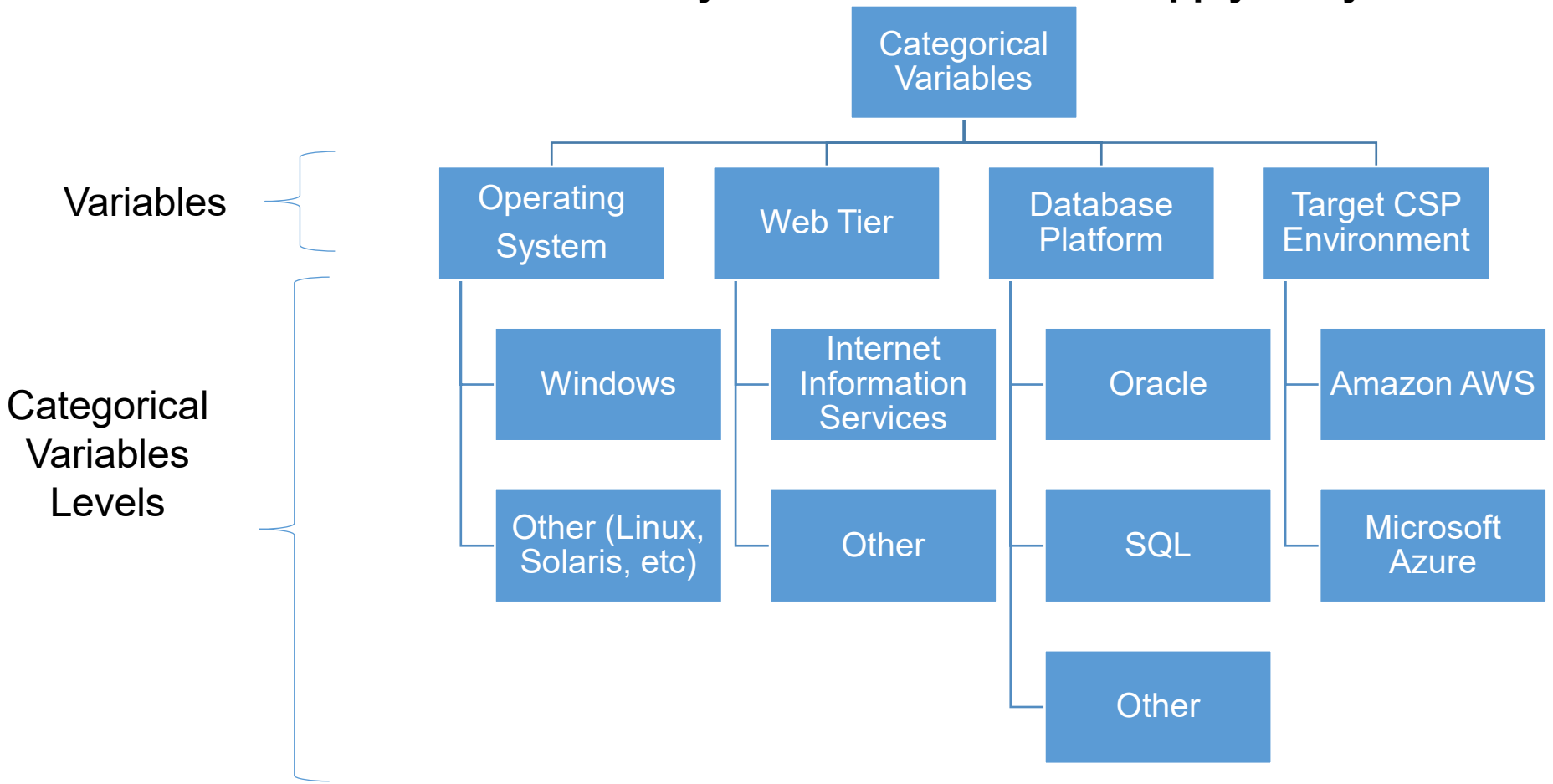
46 additional variables removed; responses too sparse or no significant correlation to migration cost

Seven (7) variables for which response was sufficiently dense; missing values among these variables led to nine (9) more applications removed from analysis

Final curated dataset: 12 applications across seven variables, 100% density

Data Cleansing and Variable Preparation

Four of the remaining seven variables were qualitative/categorical (not numerical) and needed to be converted to binary indicator variables to apply analytical methods



FY21 Model and Issues Encountered

Multiple Regression model

Migration Cost (CY20$) = $1,763,072 + 0.12044 x SLOC + (-$1,480,055) x OS_Windows + (-$1,094,075) x TargetEnvironment_AWS

Issues Encountered

- Many Windows applications migrating to AWS led to nonsensical estimates (negative cost) due to large negative coefficients in FY21 Model
- Variable selection process was not performed using a stable method or objective set of criteria
- FY21 Model required analyst intervention to produce meaningful estimates due to the above issues and relatively poor accuracy

FY22 Dataset Modifications

Many new applications have completed migrations into Cloud One; cost/technical data were available for nine of them

Introduced $SLOC^2$ as an independent variable after new model testing with non-linear and non-parametric models proved more accurate than linear parametric models

Total Registered Users removed as independent variable due to lack of data in the nine added observations

FY22 Model Testing

Needed to resolve variable selection issues

- Started with a brute force approach: test 127 combinations of predictors
- Began the arduous task of testing all linear models with hat-matrix ($\mathbf{H = X(X^TX)^{-1}X^T}$) in Excel
- Quickly realized the inefficiency of approach and began investigating alternate tools

MATLAB was selected as an alternative to Excel

- MATLAB is built to work with matrices
- MATLAB's programming language could automate tedious matrix calculations

Biggest factor: easy availability on Air Force laptops

Discovered suite of MATLAB apps for training/testing predictive models

- Trained/tested 600+ models, with cross validation, in a few minutes
- Algorithms included linear/multiple/stepwise regression, decision trees, support vector machines, and boosted and bagged trees
- **SLOC alone emerged as the best predictor variable using this approach**

Initial Testing Results

- **Initial testing proved promising**
  - **The best models tested using MATLAB's regression learner suite produced models with RMSE around $400k (CV of 74%)**
  - **That's still not great, but a significant improvement over the FY21 model's RMSE of $714k (CV of 141%) and an *enormous* improvement over averages**
  - **An interesting observation: the best models were trained using non-linear or non-parametric algorithms (ensemble trees, SVM)**

Test Statistic	Average Cost by Complexity	Average Cost by Complexity (Outliers Ignored)	FY21 Multiple Regression	Bagged Trees	Fine Gaussian SVM
Root Mean Squared Error (with LOOCV)	$10,651,400	$1,845,923	$713,760	$401,610	$412,510
Mean of Dataset (CY20$)	$700,886	$700,886	$507,663	$544,126	$544,126
Coefficient of Variation	1520%	263%	141%	74%	76%

Testing Results and Follow-up

Problems with non-linear and non-parametric models

- Gaussian SVM requires complex transformation of predictors; problematic to request of customers
- Ensemble tree methods improves accuracy at the cost of model interpretability
- Both models would require packaging as MATLAB (or Python or R) applications, likely requiring some security approval and authorization to distribute; essentially a non-starter

Since non-linear algorithms produced better models on this dataset, decided to explore a non-linear term

- Introduced a squared term: $SLOC^2$
- Driven by a desire to find the Goldilocks solution: improve accuracy, maintain model interpretability, and minimize deployment impact

FY22 Model and Statistics

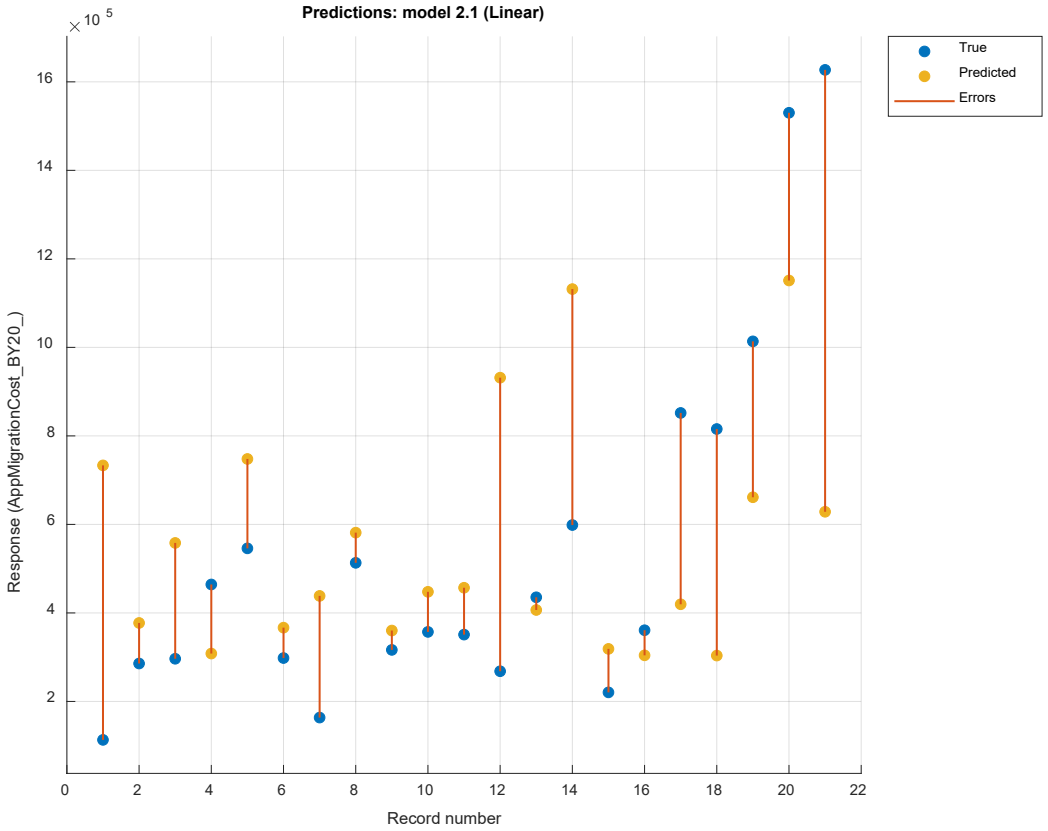
Linear Regression model

Migration Cost (CY20$) = $309,517 + 0.5386 x SLOC + (-4.3288x10^{-8}) x SLOC^2

Statistics

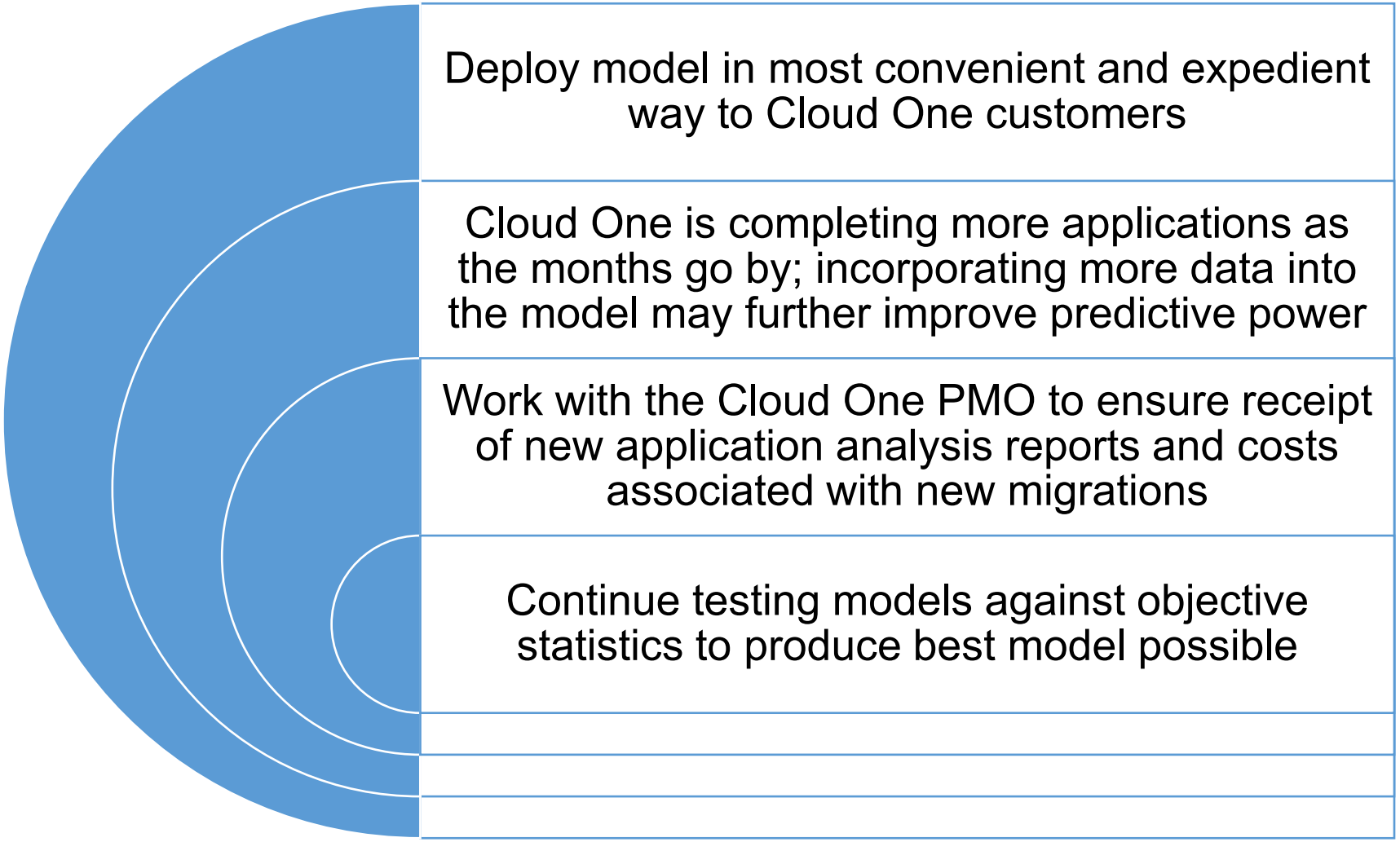
RMSE (with LOOCV): $383,250

Coefficient of Variation: 70%



A marked improvement over the FY21 model!
AND
No manual workarounds with this model!

Future Work


Deploy model in most convenient and expedient way to Cloud One customers
Cloud One is completing more applications as the months go by; incorporating more data into the model may further improve predictive power
Work with the Cloud One PMO to ensure receipt of new application analysis reports and costs associated with new migrations
Continue testing models against objective statistics to produce best model possible

Questions?

DELIVER TO COMMITMENTS

Variable List

Functional	Files - 5/W3C	Files-ASP/HTML	Files - diagram	Files - ide	Files - list	Files - pdb
Impact Level	Files - 6/bin/ant	Files - ASPX/ASCX/ASAX	Files - dic	Files - in	Files - locale	Files - pdf
Total Registered Users - Dev	Files - 6/bin/antRun	Files - bak	Files - dll	Files - ini	Files - lock	Files - Perl
Total Registered Users - Test	Files - 6/CONTRIBUTORS	Files - bat	Files - DOS Batch	Files - jar	Files - manifest	Files - pkb
Total Registered Users - Prod	Files - 6/INSTALL	Files - Bourne Shell	Files - DS_Store	Files - Java	Files - map	Files - pks
Total Registered Users - Total	Files - 6/KEYS	Files - BSD	Files - dtd	Files - java_delete	Files - Markdown	Files - pl
Avg Daily Users - Dev	Files - 6/lib/README	Files - C	Files - ear	Files - java_deleteme	Files - Master	Files - pll
Avg Daily Users - Test	Files - 6/LICENSE	Files - C#	Files - edmx	Files - java_old	Files - md	Files - plsql
Avg Daily Users - Prod	Files - 6/manual/api/package-list	Files - cache	Files - ent	Files - java_tryagainlater	Files - MF	Files - png
Avg Daily Users - Total	Files - 6/manual/LICENSE	Files - cd	Files - eot	Files-JavaScript & TypeScript	Files - mmb	Files - pom
Avg Concurrent Users - Dev	Files - 6/NOTICE	Files - class	Files - exe	Files - JavaScript	Files - mn	Files - pp
Avg Concurrent Users - Test	Files - 6/README	Files - classpath	Files - flex actionscript	Files - jbf	Files - modelproj	Files - pptx
Avg Concurrent Users - Prod	Files - 6/WHATSNEW	Files - cmd	Files - flex mxml	Files - jnlp	Files - MSBuild Script	Files - prc
Avg Concurrent Users - Total	Files - 9/LICENSE	Files - ColdFusion	Files - fmb	Files - JPG	Files - number	Files - prefix
Number of Databases	Files - aff	Files - conf	Files - fnc	Files - jshintrc	Files - nupkg	Files - prefs
Database Size (GBs)	Files - angularTree	Files - config	Files - form	Files - jsm	Files - nuspec	Files - pri
Data Volume(GB)	Files-ActionScript	Files - CopyComplete	Files - gif	Files - JSON	Files - olb	Files - project
Files - 2/LICENSE	Files - Ant	Files-C/C++ header	Files - gitignore	Files - JSP	Files - ora	Files - properties
Files - 5/Bitstream-Vera-Fonts	Files - Apache Config	Files - cshtml	Files - glsl	Files - ks	Files - original	Files - ps1
Files - 5/COLORBREWER	Files - asax	Files-C++	Files-Gencat NLS	Files - layout	Files - osql	Files - psd1
Files - 5/GPL+CP	Files - ascx	Files - csproj	Files - Gradle	Files - less	Files - otf	Files - psm1
Files - 5/LGPL	Files - ashx	Files - CSS	Files - htm	Files - lgx	Files - p7s	Files - pspimage
Files - 5/LICENSE	Files-ASP.NET	Files - db	Files - HTML	Files - lic	Files - par	Files - pubxml
Files - 5/OGC	Files - aspx	Files - defaults	Files - ico	Files - licx	Files - package-list	Files-Python

Variable List

Files - py	Files - tps	Files - XML	SLOC - 6/LICENSE	SLOC - C#	SLOC - dtd	SLOC - java_delete
Files-RDL	Files - transform	Files - xpt	SLOC - 6/manual/api/package-list	SLOC - cache	SLOC - ear	SLOC - java_deleteme
Files - rdf	Files - ts	Files - xsd	SLOC - 6/manual/LICENSE	SLOC - cd	SLOC - edmx	SLOC - java_old
Files - readme	Files - tt	Files - xsl	SLOC - 6/NOTICE	SLOC - class	SLOC - ent	SLOC - java_tryagainlater
Files - resources	Files - ttf	Files - xslt	SLOC - 6/README	SLOC - classpath	SLOC - eot	SLOC-JavaScript & TypeScript
Files-Ruby	Files - ttinclude	Files - xul	SLOC - 6/WHATSNEW	SLOC - cmd	SLOC - exe	SLOC - JavaScript
Files - resx	Files - txt	Files - XULRunner	SLOC - 9/LICENSE	SLOC - ColdFusion	SLOC - flex actionscript	SLOC - jbf
Files - Saas	Files - uml	Files - yml	SLOC - aff	SLOC - conf	SLOC - flex mxml	SLOC - jnlp
Files - Sass	Files-Visual Basic	Files - zip	SLOC - angularTree	SLOC - config	SLOC - fmb	SLOC - JPG
Files - scss	Files - vbs	Files - Total	SLOC - Ant	SLOC - CopyComplete	SLOC - fnc	SLOC - jshintrc
Files - sequencediagram	Files - Visualforce	SLOC - 2/LICENSE	SLOC - Apache Config	SlOC-C/C++ header	SLOC - form	SLOC - jsm
Files - settings	Files - vspscc	SLOC - 5/Bitstream-Vera-Fonts	SLOC - asax	SLOC-C++	SLOC - gif	SLOC - JSON
Files - sh	Files - vssscc	SLOC - 5/COLORBREWER	SLOC - ascx	SLOC - CSS	SLOC - gitignore	SLOC - JSP
Files - Shell	Files - vsx	SLOC - 5/GPL+CP	SLOC - ashx	SLOC - cshtml	SLOC-Gencat NLS	SLOC - ks
Files-Silverlight (XMAP/XAP)	Files - Vue	SLOC - 5/LGPL	SLOC - aspx	SLOC - csproj	SLOC - glsl	SLOC - layout
Files - skin	Files - war	SLOC - 5/LICENSE	SLOC-ActionScript	SLOC - CSS	SLOC - Gradle	SLOC - less
Files - sln	Files - wav	SLOC - 5/OGC	SLOC-ASP.NET	SLOC - db	SLOC - htm	SLOC - lgx
Files - SQL	Files-Windows Resource File	SLOC - 5/W3C	SLOC-ASP/HTML	SLOC - defaults	SLOC - HTML	SLOC - lic
Files - suffix	Files - woff	SLOC - 6/bin/ant	SLOC - ASPX/ASCX/ASAX	SLOC - diagram	SLOC - ico	SLOC - licx
Files - svg	Files - woff2	SLOC - 6/bin/antRun	SLOC - bak	SLOC - dic	SLOC - ide	SLOC - list
Files - swf	Files - xaml	SLOC - 6/CONTRIBUTORS	SLOC - bat	SLOC - dll	SLOC - in	SLOC - locale
Files - targets	Files - xfdl	SLOC - 6/INSTALL	SLOC - Bourne Shell	SLOC-DOS	SLOC - ini	SLOC - lock
Files - text	Files - xhtml	SLOC - 6/KEYS	SLOC - BSD	SLOC - DOS Batch	SLOC - jar	SLOC - manifest
Files - tpignore	Files - xlsx	SLOC - 6/lib/README	SLOC - C	SLOC - DS_Store	SLOC - Java	SLOC - map

Variable List

SLOC - Markdown	SLOC - pl	SLOC - resx	SLOC - ttinclude	SLOC - xslt	Web Server CPU - Prod
SLOC - Master	SLOC - pll	SLOC-Ruby	SLOC - txt	SLOC - xul	Web Server RAM (GB) - Prod
SLOC - md	SLOC - plsql	SLOC - Saas	SLOC - uml	SLOC - XULRunner	Web Server Disk (GB) - Prod
SLOC - MF	SLOC - png	SLOC - Sass	SLOC - vbs	SLOC - yml	Web Server CPU - COOP
SLOC - mmb	SLOC - pom	SLOC - scss	SLOC - Visualforce	SLOC - zip	Web Server RAM (GB) - COOP
SLOC - mn	SLOC - pp	SLOC - sequencediagram	SLOC - vspscc	SLOC - Total	Web Server Disk (GB) - COOP
SLOC - modelproj	SLOC - pptx	SLOC - settings	SLOC-Visual Basic	Database Server CPU	User Locations
SLOC - MSBuild Script	SLOC - prc	SLOC - sh	SLOC - vssscc	Database Server RAM (GB) - Dev	Environments
SLOC - number	SLOC - prefix	SLOC - Shell	SLOC - vsx	Database Server RAM (GB) - Pre-Prod	Operating System
SLOC - nupkg	SLOC - prefs	SLOC-Silverlight (XMAP/XAP)	SLOC - Vue	Database Server RAM (GB) - Prod	Web Tier
SLOC - nuspec	SLOC - pri	SLOC - skin	SLOC - war	Database Server Disk (GB) Dev	Database Platform
SLOC - olb	SLOC - project	SLOC - sln	SLOC-Windows Resource File	Database Server Disk (GB) - Pre-Prod	Inbound Interfaces
SLOC - ora	SLOC - properties	SLOC - SQL	SLOC - wav	Database Server Disk (GB) - Prod	Inbound Interface Frequency
SLOC - original	SLOC - ps1	SLOC - suffix	SLOC - woff	External Database Connection?	Outbound Interfaces
SLOC - osql	SLOC - psd1	SLOC - svg	SLOC - woff2	Servers	Outbound Interface Frequency
SLOC - otf	SLOC - psm1	SLOC - swf	SLOC - xaml	Number of Servers	Total Interfaces
SLOC - p7s	SLOC - pspimage	SLOC - targets	SLOC - xfdl	Server Locations	
SLOC - package-list	SLOC - pubxml	SLOC - text	SLOC-XML	Server Locations (Add'l)	
SLOC - par	SLOC-Python	SLOC - tpignore	SLOC - xhtml	Web Server CPU - Dev/Test	
SLOC - pdb	SLOC - py	SLOC - tps	SLOC - xlsx	Web Server RAM (GB) - Dev/Test	
SLOC - pdf	SLOC - readme	SLOC - transform	SLOC - XML	Web Server Disk (GB) - Dev/Test	
SLOC - Perl	SLOC - rdf	SLOC - ts	SLOC - xpt	Web Server CPU - PreProd	
SLOC - pkb	SLOC-RDL	SLOC - tt	SLOC - xsd	Web Server RAM (GB) - PreProd	
SLOC - pks	SLOC - resources	SLOC - ttf	SLOC - xsl	Web Server Disk (GB) - PreProd	

Data Cleansing, Variable Preparation and Normalization (extra detail)

- 46 additional variables removed
  - Functional Customer (1) – no correlation found to App Migration cost
  - Impact level (1) – too much uniformity in responses, 82% of applications in dataset are IL4
  - Average Daily Users and Average Concurrent Users (11) – too many empty entries
  - Database Size and Data Volume (2) – too many empty entries
  - Total Interfaces (5) – too many empty entries
  - Operating Environment System Specifications (26) – too many empty entries
- Resulted in seven (7) final variables for which response was sufficiently dense
- There were still missing values among these seven (7) final variables, so nine (9) more applications had to be removed from the analysis
- 15 applications either started migration and didn't finish or didn't start migration at all after app analysis, had to remove these as well
- Final curated dataset included 12 applications across seven variables at 100% response density
  - Three quantitative variables: Total Registered Users, Total Files, Total SLOC
  - Four qualitative/categorical variables: Current Operating System, Current Web Tier, Current Database Platform, and Target CSP Environment