Predicting Startup Status

William Fang

23/06/2020

Contents

1 Introduction

The aim of this report is to develop a classification model to predict the status of a startup, based primarily on the funding of the company. The data set used is "StartUp Investments (Crunchbase)", available from Kaggle. It provides the status of a company (acquired, operating, or closed), funding and other feature data.

Data was first cleaned, then explored and analysed to allow selection of features and models to apply. The data was split into training and validation sets, then models trained and evaluated against these sets.

The data set can be downloaded as a zipped CSV file, and has been added to the github repository of this project. The uncompressed file is approximately 12Mb, and there are 49176 rows, and 39 variables.

1.1 Data Features

The variable names of the data frame are as below:

```
 [1] "permalink"           "name"                "homepage_url"
 [4] "category_list"       "market"              "funding_total_usd"
 [7] "status"              "country_code"        "state_code"
[10] "region"              "city"                "funding_rounds"
[13] "founded_at"          "founded_month"       "founded_quarter"
[16] "founded_year"        "first_funding_at"    "last_funding_at"
[19] "seed"                "venture"             "equity_crowdfunding"
[22] "undisclosed"         "convertible_note"    "debt_financing"
```

```
[25] "angel"                "grant"                "private_equity"
[28] "post_ipo_equity"      "post_ipo_debt"        "secondary_market"
[31] "product_crowdfunding" "round_A"              "round_B"
[34] "round_C"              "round_D"              "round_E"
[37] "round_F"              "round_G"              "round_H"
```

The data set has no meta data explaining the definitions of the features, and instead reasonable interpretations are assumed.

1.1.1 Market and Categories

The `market` variable is a string representing the main market the startup is targeting. `category_list` contains one or more categories the startup belongs to. Each category is separated by a |, and there is no specific ordering to the list.

Table 1: Sample category_list Values

x
|Entertainment|Politics|Social Media|News|
|Games|
|Publishing|Education|
|Electronics|Guides|Coffee|Restaurants|Music|iPhone|Apps|Mobile|iOS|E-Commerce|
|Tourism|Entertainment|Games|
|Software|

1.1.2 Location

There are four variables that describe location. `country_code` which is a 3 character string, `state_code` a 2 character string for companies within the US, `region` and `city` which are strings.

1.1.3 Status

The focus of this report is to create a model for predicting status. The variable `status` has the following values:

Table 2: Status Values

x
acquired
operating
NA
closed

There are companies with a missing status value, these rows will be removed as they can't be used.

1.1.4 Date

The following features represents date related data, and are self explanatory: `founded_at`, `founded_month`, `founded_quarter`, `founded_year`, `first_funding_at`, `last_funding_at`

1.1.5 Funding

The variable `funding_total_usd` represents total funding in US dollars, `funding_rounds` how many rounds (round A, B etc), and the rest represent the total broken down by type (angel investor, venture capital etc).

1.1.6 Ignored Variables

The following variables will be ignored, as they are very unlikely to have any predictive power. `permalink`, which is a hyperlink from the Techcrunch data, and `homepage_url`, which is the company's web page.

The variable `name`, may be some sort of predictive power here. For example, a catchy or memorable startup name might attract more attention and lead to more investment, consumer interest etc. However in terms of this dataset it isn't likely to be useful.

2 Analysis

2.1 Data Cleaning

The first step is to examine the data, looking for missing data and bad/invalid values, and wrong data types.

2.1.1 Missing Values

The following table shows the counts of missing (NA) values in the data set:

Table 3: Variables With Missing Values Counts

variable	count
state_code	19172
founded_month	10909
founded_quarter	10909
founded_year	10909
founded_at	10838
city	6088
country_code	5247
region	5247
market	3959
category_list	3951
homepage_url	3441
status	1310
angel	1
convertible_note	1
debt_financing	1
equity_crowdfunding	1
first_funding_at	1
funding_rounds	1
funding_total_usd	1
grant	1
last_funding_at	1
post_ipo_debt	1
post_ipo_equity	1

variable	count
private_equity	1
product_crowdfunding	1
round_A	1
round_B	1
round_C	1
round_D	1
round_E	1
round_F	1
round_G	1
round_H	1
secondary_market	1
seed	1
undisclosed	1
venture	1

There are a few variables where there is only 1 NA value, and it turns out they all belong to one entry which can be removed.

The following plot shows the percentage of a variable that has NA values (code adapted from Laufer (n.d.)):

Handling of NA values for the rest of the variables will be discussed in the following sections.

2.1.2 Status

Since the aim is to predict the status of a company, the rows that don't have a valid status can't be used and are simply removed. The data type is changed from character to factor.

2.1.3 Unused Variables

The features `name`, `permalink`, and `homepage_url` are unlikely to have predictive power. However it will be useful to have a unique ID to identify rows. Examining the names shows there are actually 74 names that occur more than once. Some of them are due to duplicated/erroneous data, while others are due to different companies having the same name. Cleaning of bad rows is done manually by looking at each set of duplicate rows. From a first look at the duplicate rows, it is common for the `homepage_url` to be the same, so this was used to help reduce the number of rows that needed to be manually inspected.

In light of names not being unique, and not all rows having a `homepage_url`, `permalink` will be used as the unique ID. However checking `permalink` for duplicates also shows one pair with the same value, which also is removed. The `name` and `homepage_url` will not be removed yet as they may help with cleaning of other features.

2.1.4 Market and Categories

There are 752 unique values for market, including NA values which will be replaced with the string "Unknown", and the type converted to factor.

NA values for `category_list` will be treated the same as for `market` by replacing them with "Unknown". Looking at the data shows there are 6 rows where there is truncation, however these will not be changed since there are only a few.

The `category_list` field needs to be split into individual categories to look for status - category relationships. There are 824 categories.

2.1.5 Date

The columns `founded_year`, `founded_quarter`, and `founded_month` should have been derived from `founded_at`, so these will be removed and recreated if needed from the cleaned `founded_at` variable.

Table 4: Min/Max Dates

min_founded	max_founded	min_first_fund	max_first_fund	min_last_fund	max_last_fund
1636-09-08	2014-12-13	1-05-14	2014-12-24	1-05-14	2015-01-01

There are "startups" that were founded in the 1600's. Companies that old are not likely to be considered startups. Somewhat arbitrarily, restrict the range to companies founded in or after the 1990's.

It can be seen that `first_funding_at` and `last_funding_at` have invalid values, which need correction.

Table 5: Bad Date Format first_funding_at

permalink	founded_at	first_funding_at	last_funding_at
/organization/agflow	2012-08-01	20-06-14	2013-06-01
/organization/buru-buru	2012-01-01	19-11-20	2013-04-01
/organization/exploco	2014-10-01	201-01-01	201-01-01
/organization/nubank	2013-01-01	7-05-13	2014-09-25
/organization/peoplegoal	2014-01-10	1-05-14	1-05-14
/organization/securenet-payment-systems	1997-01-01	11-11-14	2012-07-24

Table 6: Bad Date Format last_funding_at

permalink	founded_at	first_funding_at	last_funding_at
/organization/exploco	2014-10-01	201-01-01	201-01-01
/organization/peoplegoal	2014-01-10	1-05-14	1-05-14

These show 10 rows that will manually be corrected.

The funding at variables have no missing values, while `founded_at` does. A reasonable value to use in place of misssing values is `first_funding_at`.

There are no values where the `last_funding_at` date is earlier than the `first_funding_at` date. However there are 7.3% of rows where the funding date is earlier than the founding date.

Table 7: Funded Earlier Than Founded

name	founded_at	first_funding_at	last_funding_at	tdiff
Jumper Networks	2008-01-05	1960-01-01	1960-01-01	-17536 days
AndrewBurnett.com Ltd	2008-10-01	1974-01-01	1974-01-01	-12692 days
UTStarcom	2010-09-01	1996-12-01	2010-02-01	-5022 days
ActionBase	2007-03-01	1994-07-09	1994-07-09	-4618 days
AVI Web Solutions Pvt. Ltd.	2010-01-01	1998-07-14	2008-02-06	-4189 days
Leyou software	2012-04-01	2000-12-01	2008-06-01	-4139 days
imo.im	2007-04-01	1995-12-12	2013-05-23	-4128 days
SuperData Research	2009-08-01	1999-04-01	2010-07-01	-3775 days
iLike	2012-04-28	2002-01-01	2006-01-01	-3770 days
Dobleas	2012-06-06	2002-02-22	2013-03-07	-3757 days

The above table shows a sample of the rows, ordered with the biggest descrepency first. The first 2 rows can be manually corrected by using the founding date, while the rest can be changed so that the founding date is the first funded value.

The exact date is unlikely to be useful for prediction, a less precise value like year will generalize better, so all dates are converted to just the year.

2.1.6 Location

The location related features are `country_code`, `state_code`, `region` and `city`, and all have missing values.

Table 8: NA Percentages For Location Features

country_code	state_code	region	city
7.9	35.8	7.9	9

The value "ZZZ" will be used to indicate missing country codes, missing regions and city will use "Unknown". State code is only applicable to US companies and will be ignored, instead of putting the rest of the world into a single state code. All will be turned into factors.

2.1.7 Funding

The `funding_total_usd` variable has been read in as type character. As an example:

Table 9: Funding Sample

name	funding_total_usd	seed	venture
#waywire	17,50,000	1750000	0

From this it can be seen that the correct intepretation of 17,50,000 is the same as the value in the `seed` column i.e. 1750000. Converting involves stripping the comma, then converting to numeric type.

In addition, there are values consisting of a single dash character indicating a missing value. Rows with these values will be dropped.

The rest of the variables are numeric and require no cleaning.

2.2 Data Exploration

2.2.1 Status

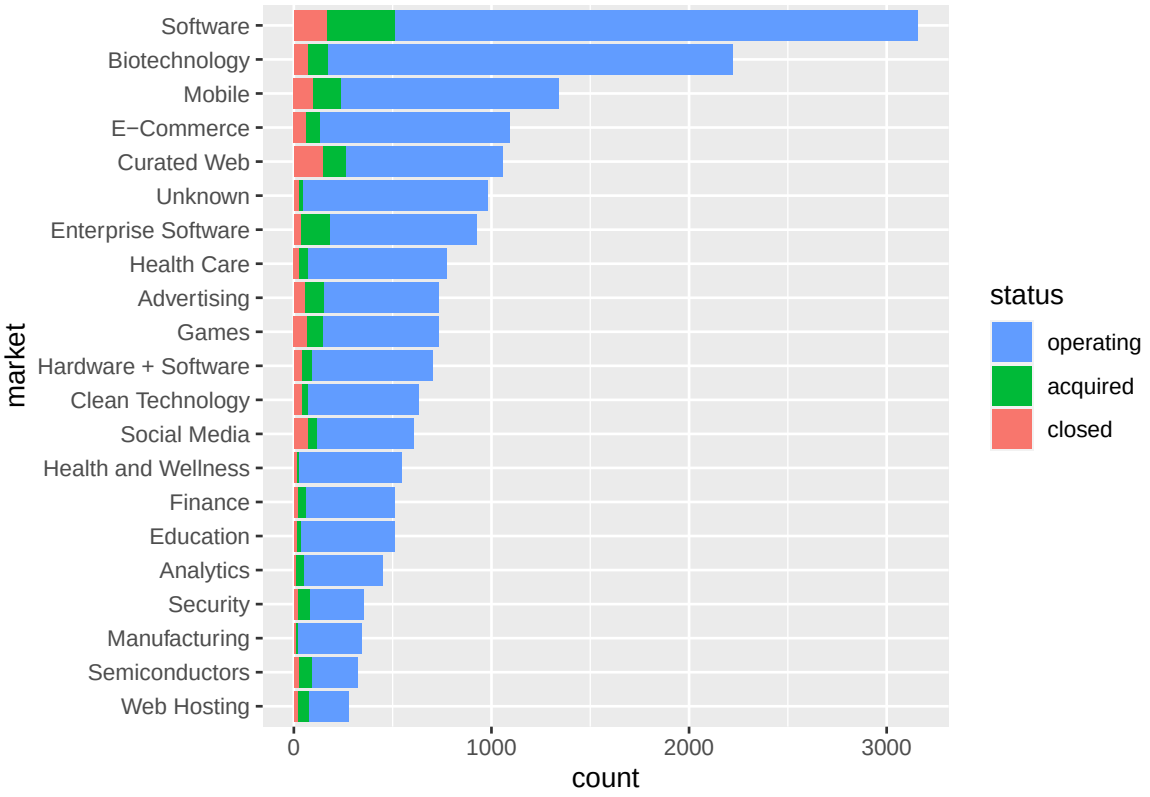
The three status values have the following counts:

Table 10: Status Counts

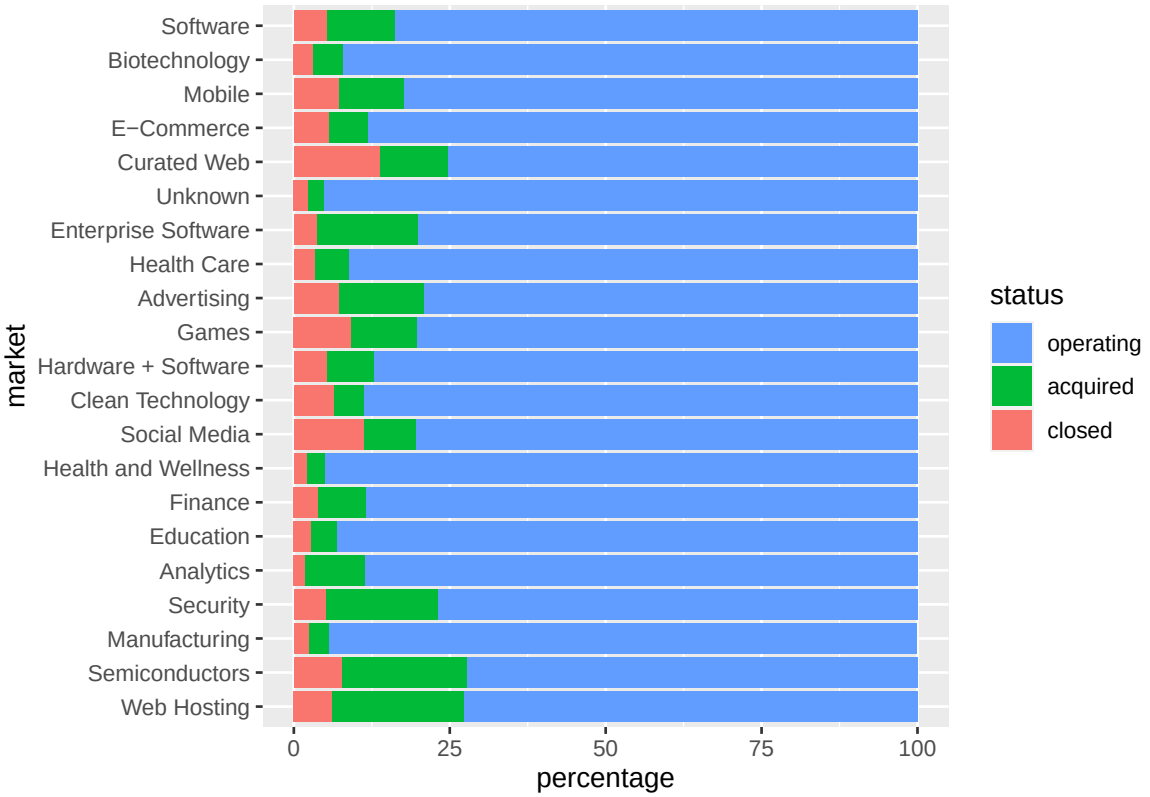
status	count	percentage
operating	26305	86.4
acquired	2520	8.3
closed	1628	5.3

2.2.2 Market and Categories

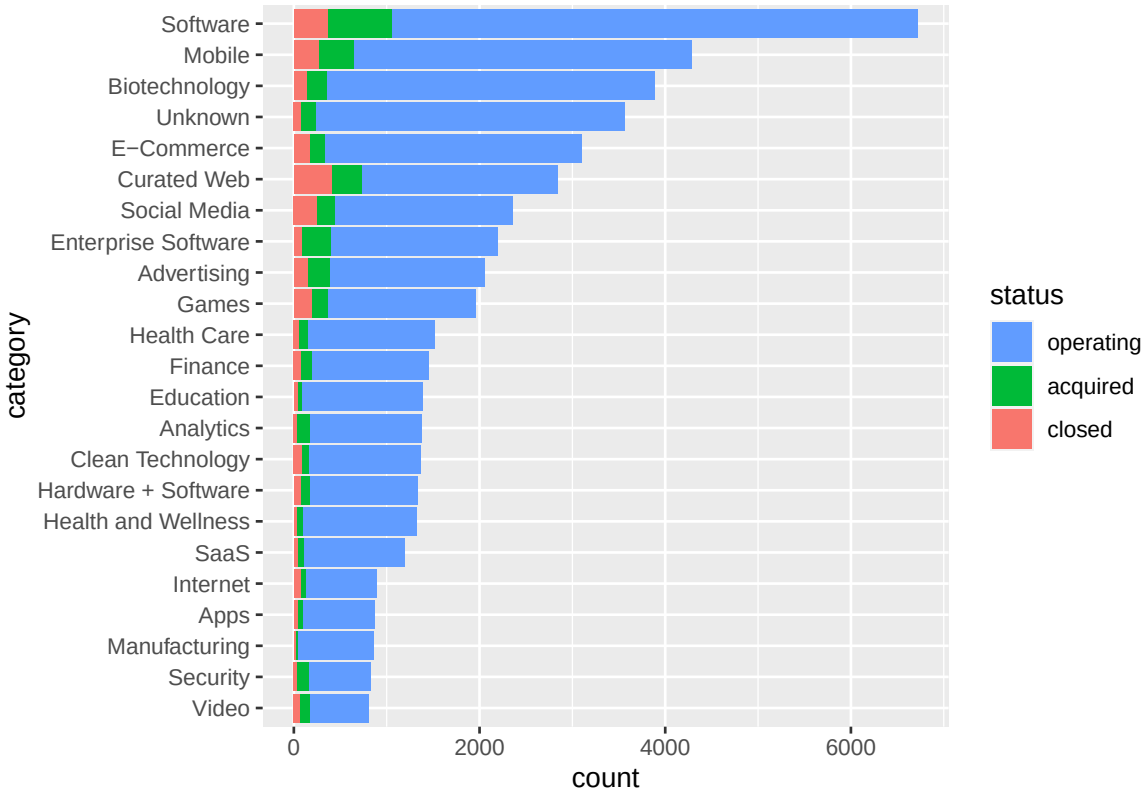
The following histogram plot shows the counts of the status per market (filtered for markets where there are at least 250 companies).

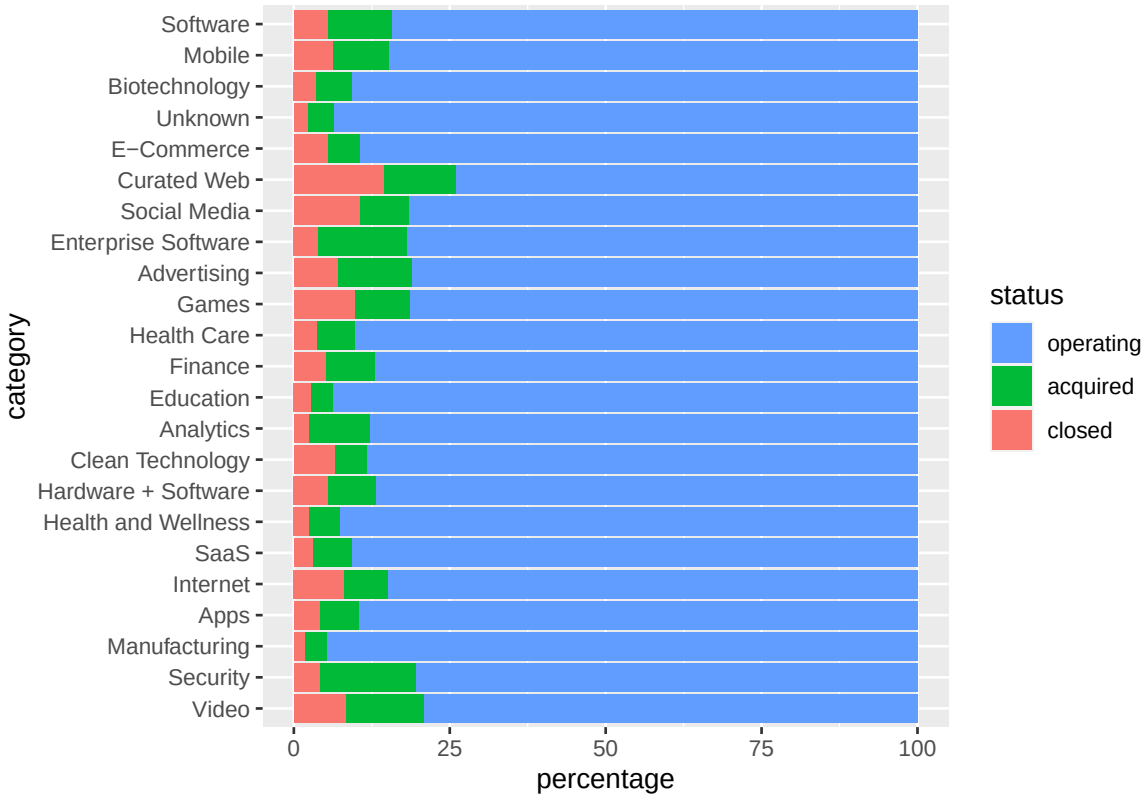


The above plot shows the proportion of statuses are not uniform and vary by market type. This is better visualized by plotting percentages of each status per market:



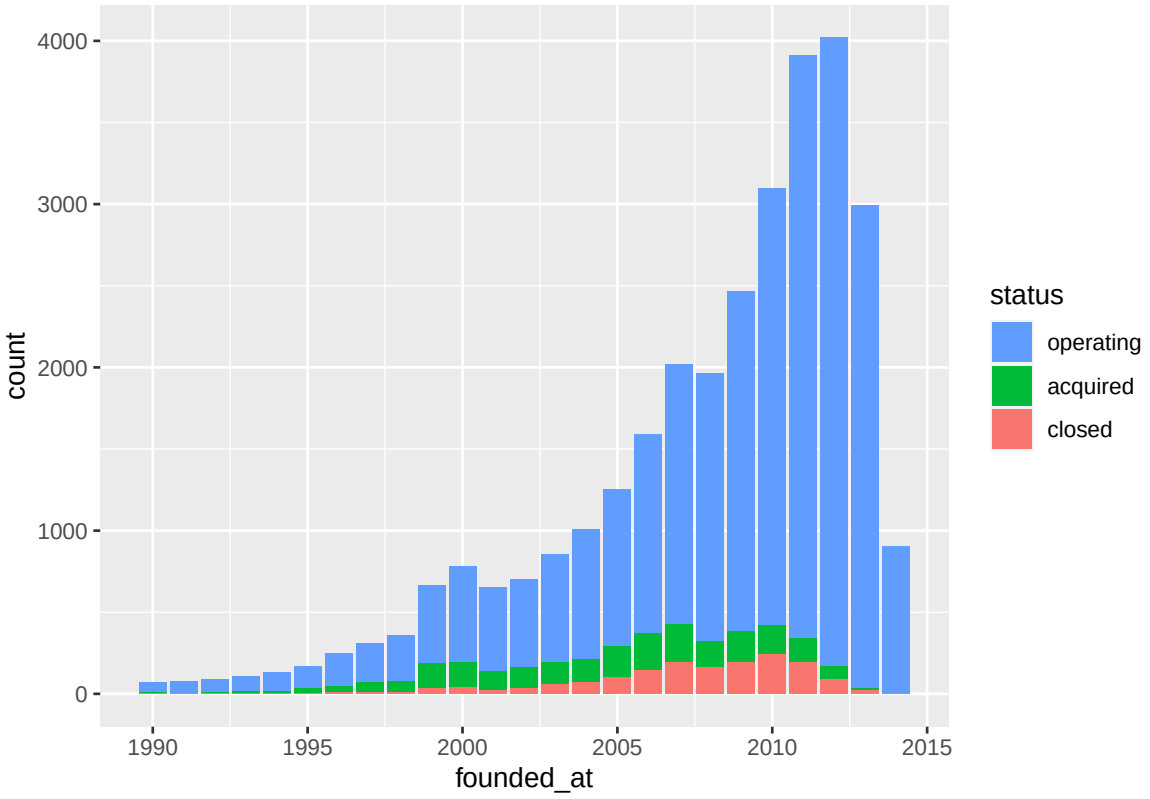
Creating plots for categories similar to those for `market`:

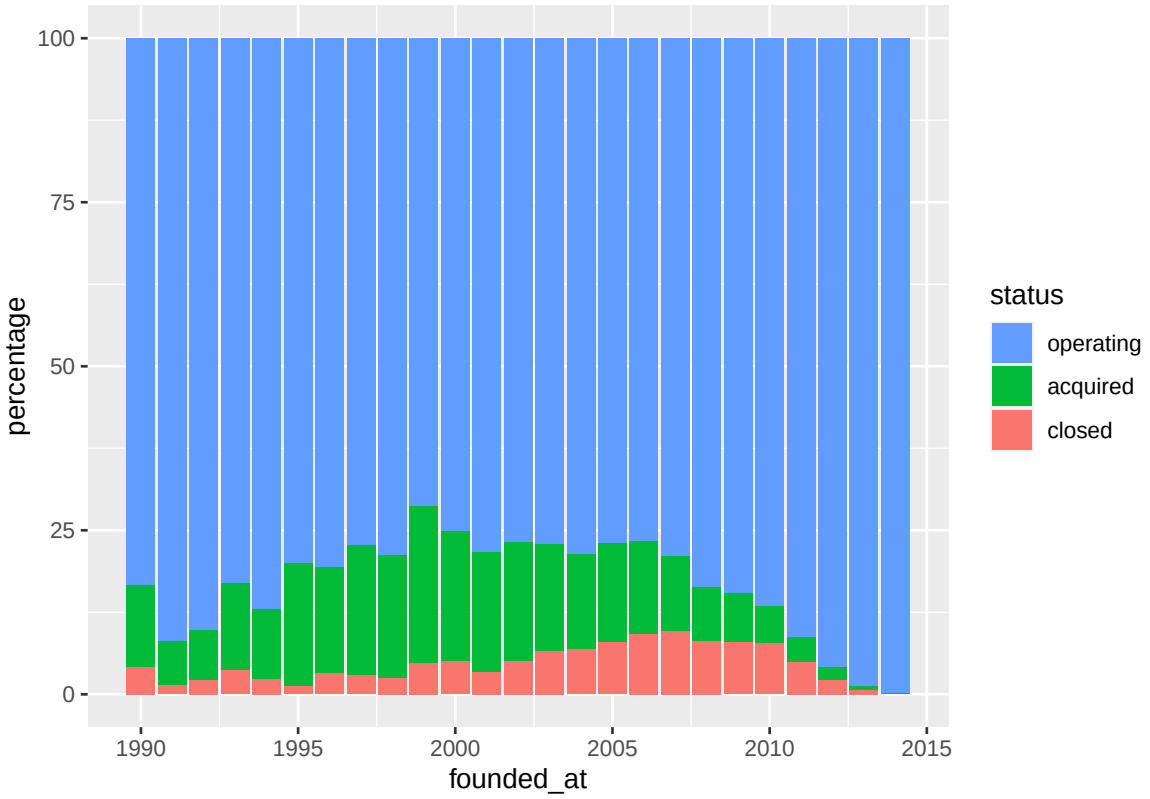




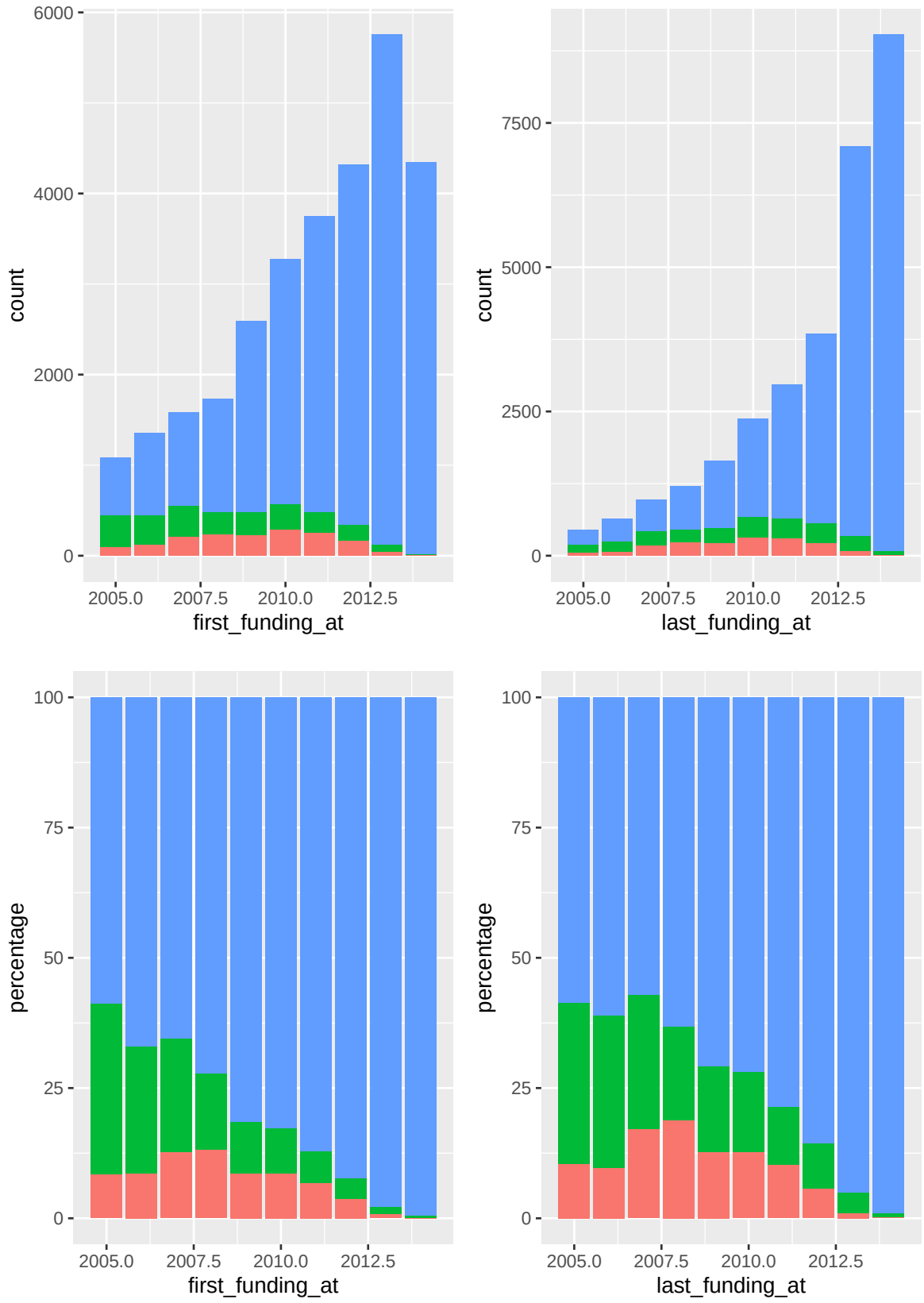
2.2.3 Date

Plotting the status vs years:



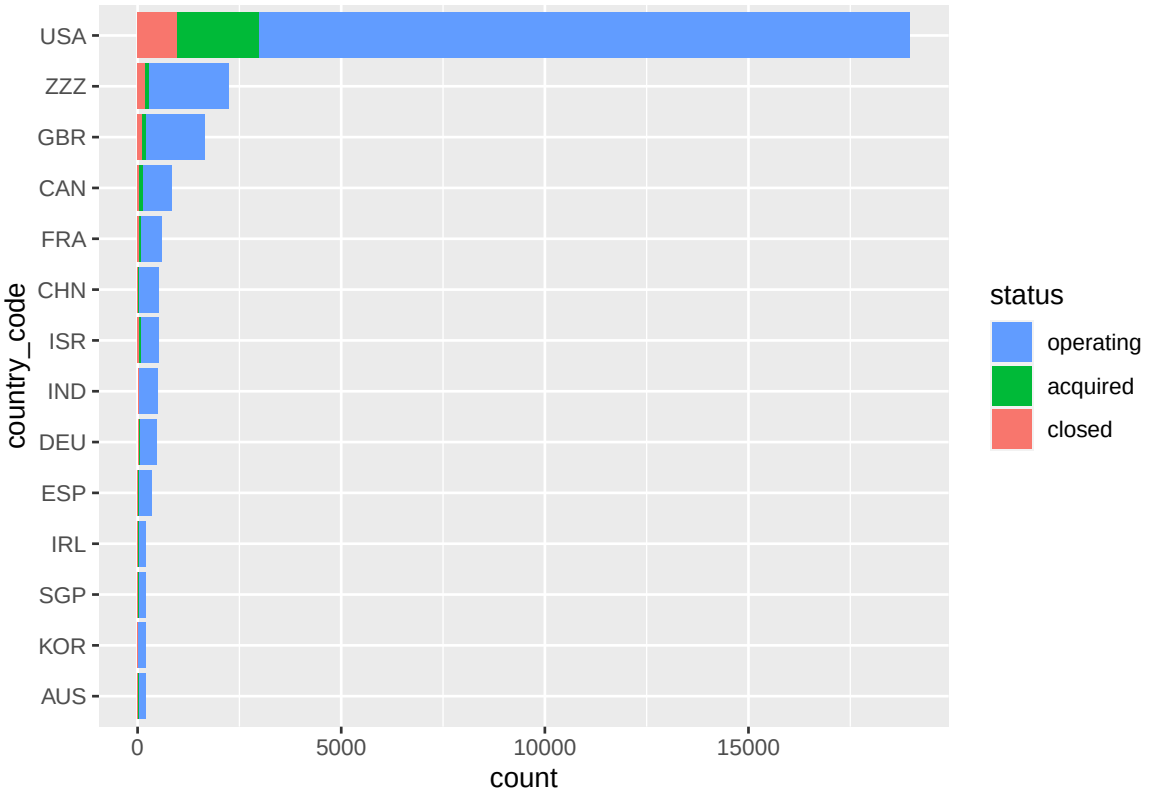


There is definite variation of the status proportions through the years. Creating similar plots for `first_funding_at` and `last_funding_at` shows variation across years as well (filtering out years with less than 250 rows):

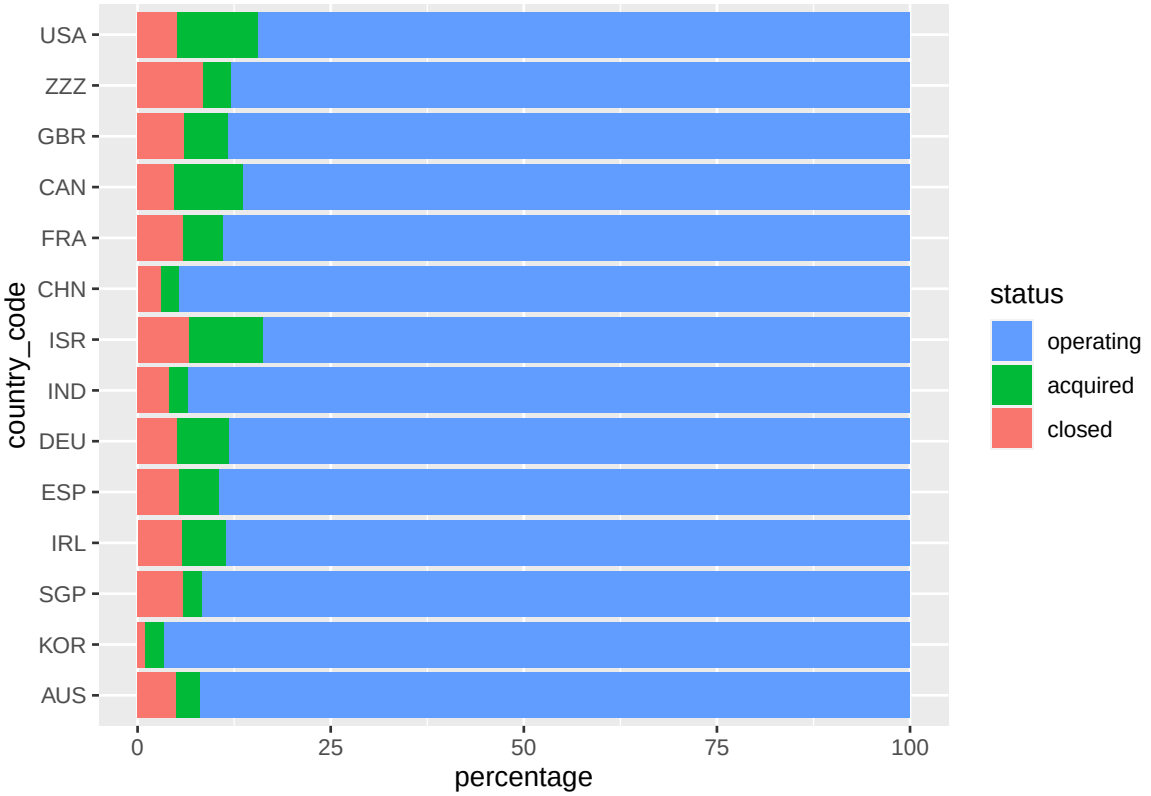


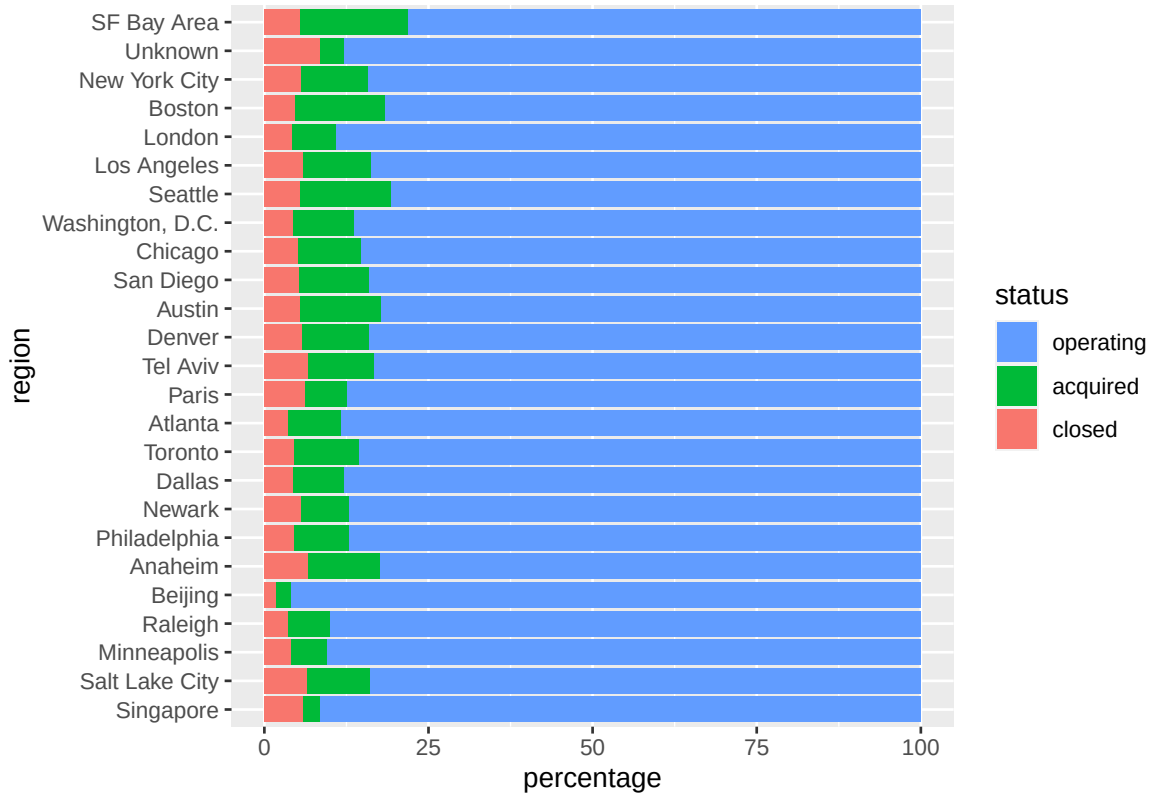
2.2.4 Location

Create plots of status versus `country_code`, `region`, and `city`, small sample sizes are not plotted.

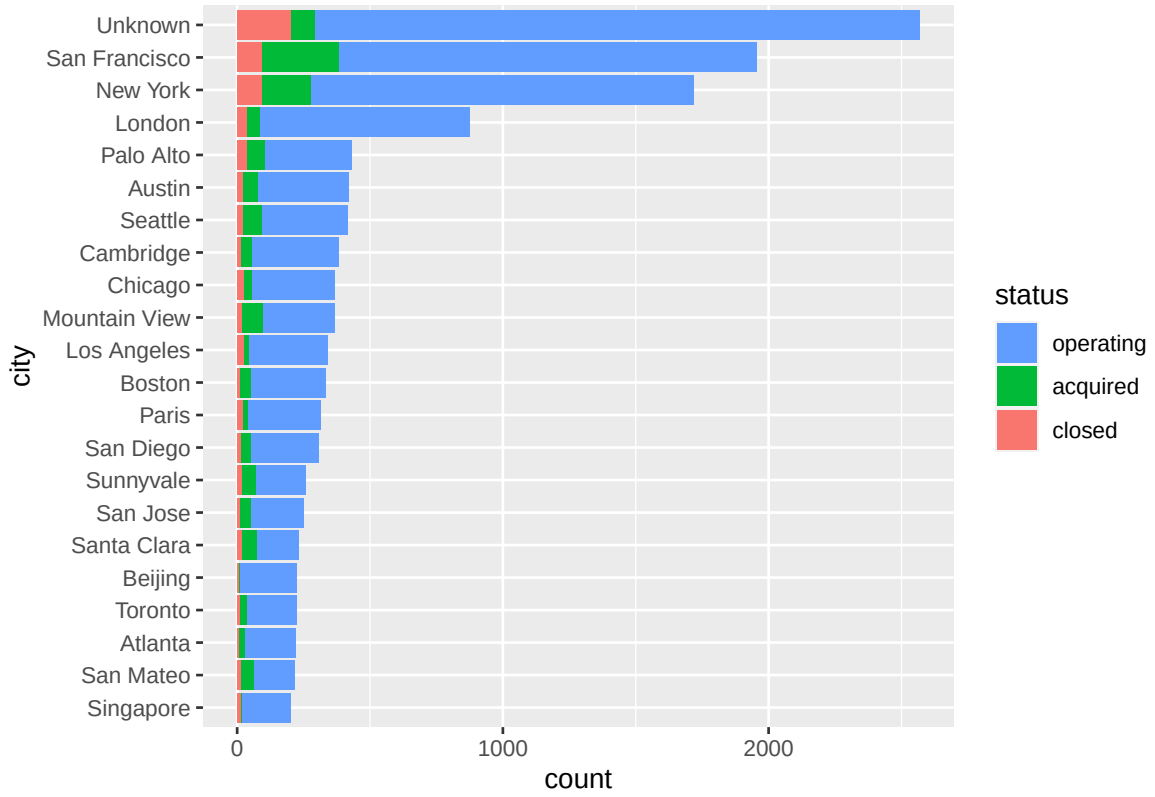


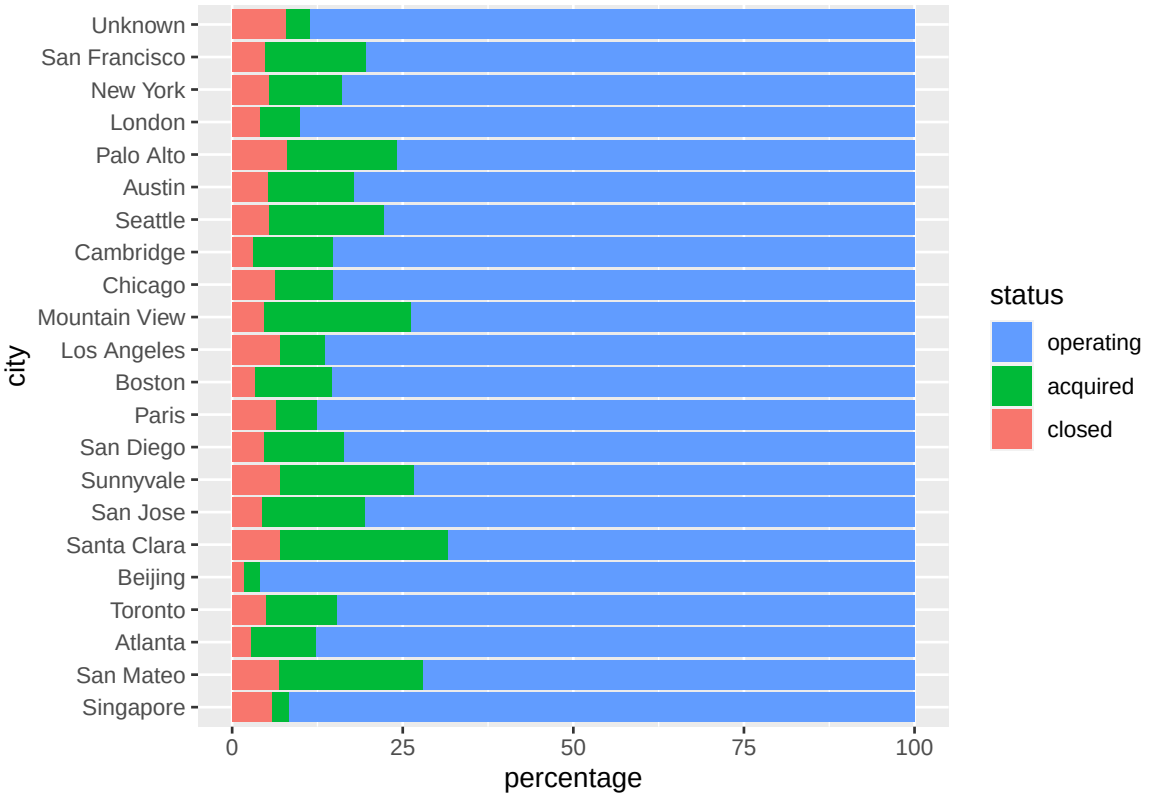
Here it can be seen that the number of US startups is significantly greater than those from other countries.





At a regional level, there is stronger variation in the percentages compared to the country level.

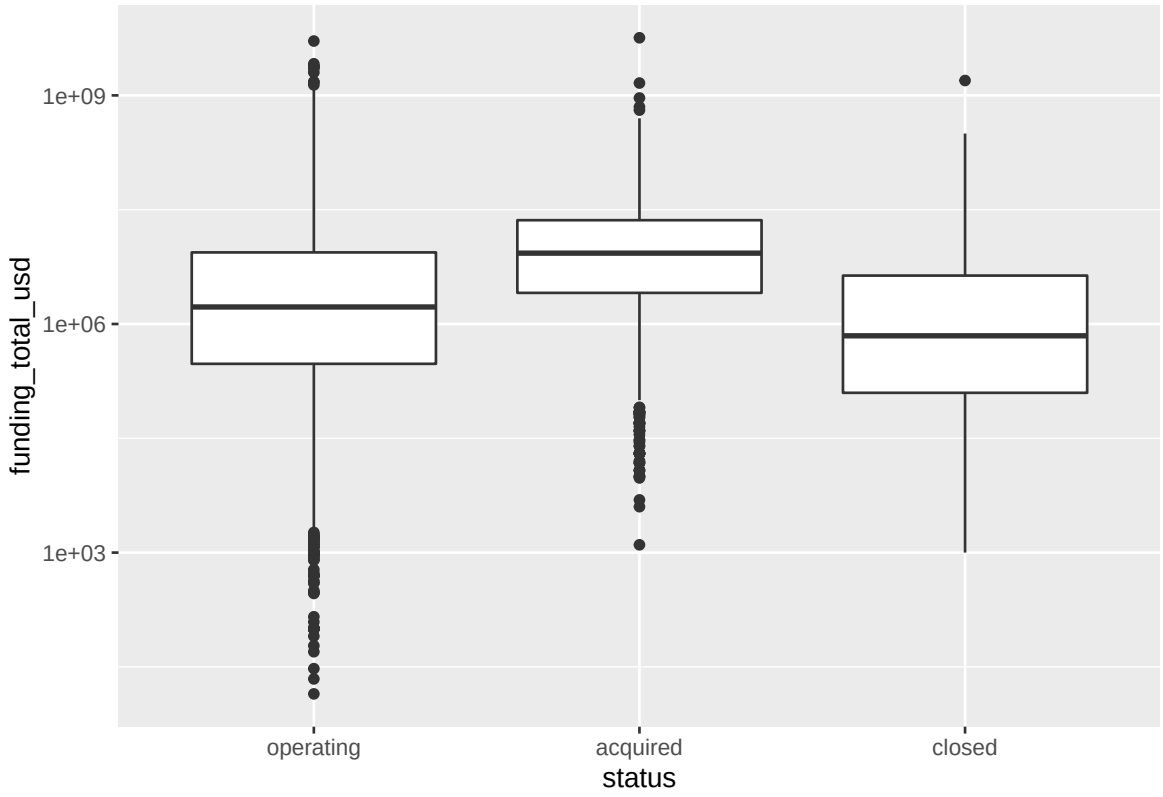




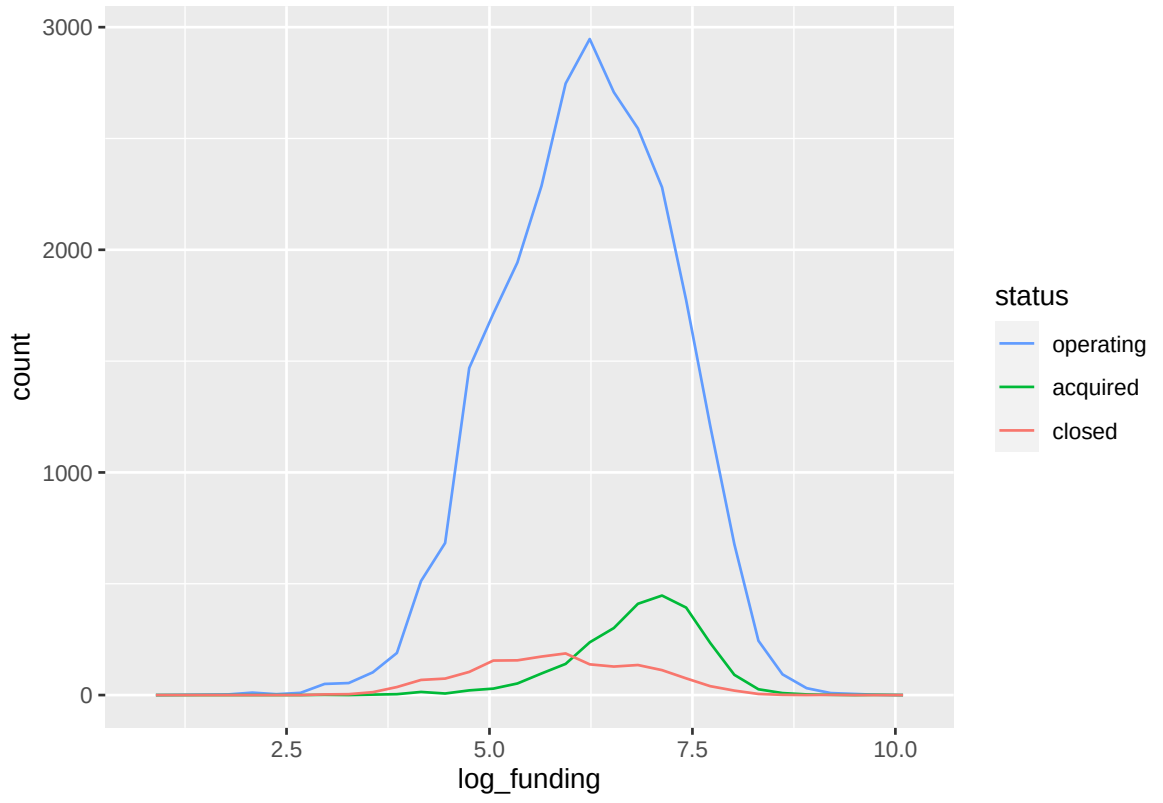
At the city level there is also stronger variation than at the region, but using this is likely too fine grained and lead to overfitting.

2.2.5 Funding

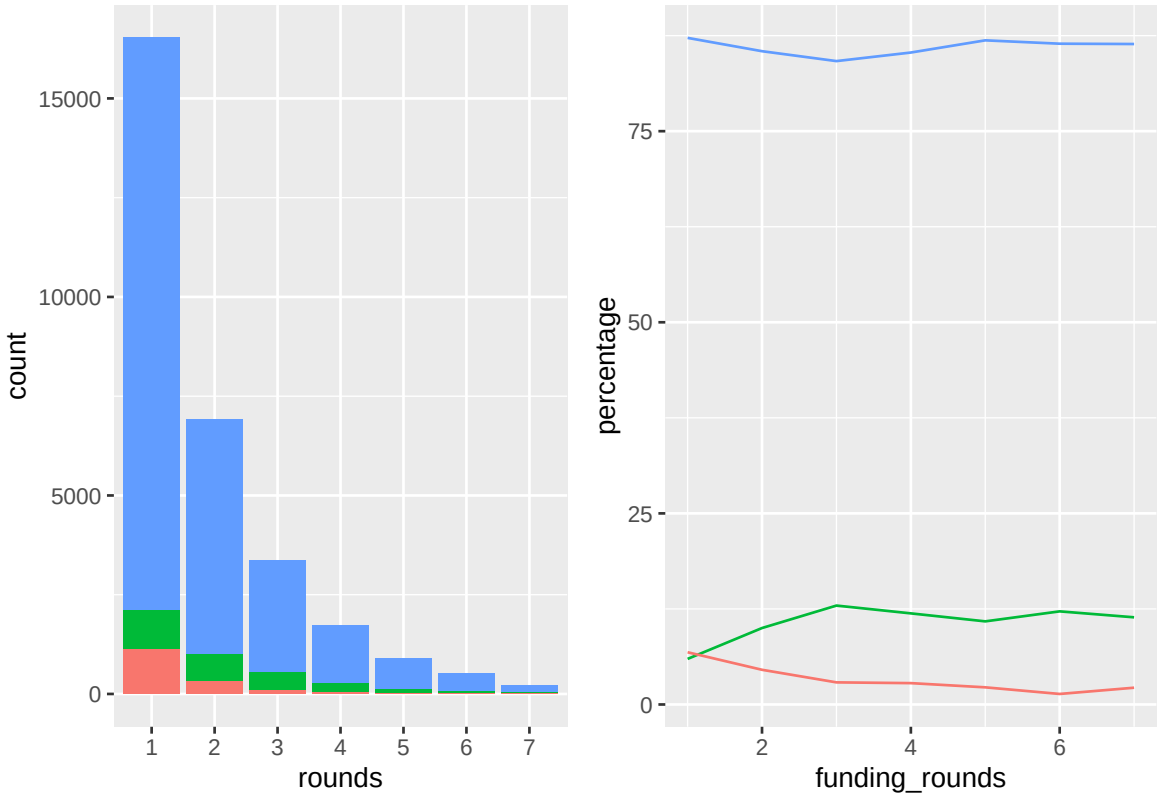
Box plot of funding against status shows startups with acquired status tend to have more funding, and those with closed status slightly less funding.



A frequency plot of log10 of the funding shows that the operating status dominates across all funding values, but there are still ranges where there is an increased chance of a company having the closed or acquired status.



Plotting status frequencies/proportions against number of funding rounds shows having more funding rounds correlates to a higher proprtion of companies being in the acquired state. The reverse is true for being in the closed state, with more rounds equating to less chance of being in the closed state.



The summary statistics of the remaining features shows that the majority of values are 0, as an example:

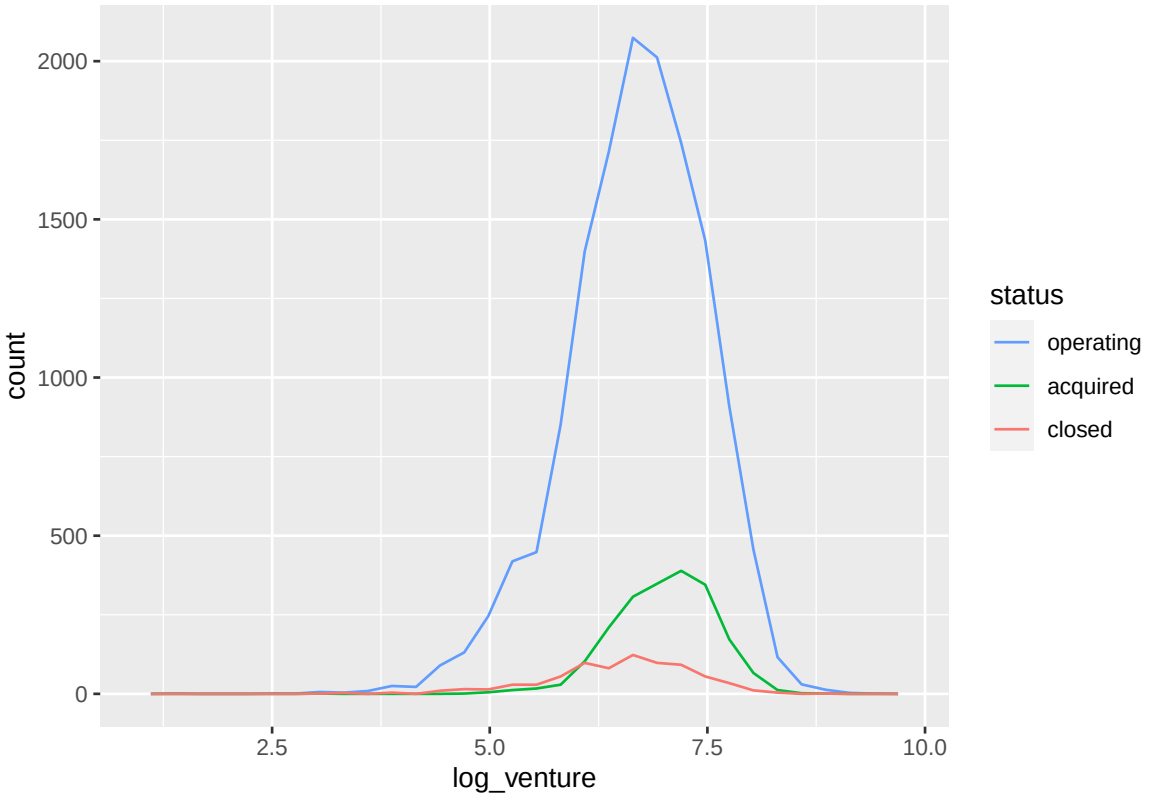
Table 11: Example Summary Statistics Funding Types

venture	seed	angel	round_A	private_equity
Min. :0.000e+00	Min. : 0	Min. : 0	Min. : 0	Min. :0.000e+00
1st Qu.:0.000e+00	1st Qu.: 0	1st Qu.: 0	1st Qu.: 0	1st Qu.:0.000e+00
Median :5.000e+05	Median : 0	Median : 0	Median : 0	Median :0.000e+00
Mean :9.857e+06	Mean : 298251	Mean : 88849	Mean : 1594489	Mean :2.351e+06
3rd Qu.:7.500e+06	3rd Qu.: 137584	3rd Qu.: 0	3rd Qu.: 0	3rd Qu.:0.000e+00
Max. :2.351e+09	Max. :100000000	Max. :43923865	Max. :225000000	Max. :2.600e+09

Only the `venture` field has the possibility of being useful. As confirmed by checking using `nearZeroVar()`, the following variables will be ignored for prediction:

```
 [1] "seed"                 "equity_crowdfunding"  "undisclosed"
 [4] "convertible_note"     "debt_financing"       "angel"
 [7] "grant"                "private_equity"       "post_ipo_equity"
[10] "post_ipo_debt"        "secondary_market"     "product_crowdfunding"
[13] "round_A"              "round_B"              "round_C"
[16] "round_D"              "round_E"              "round_F"
[19] "round_G"              "round_H"
```

Plot of counts versus log10 of venture funding amount has a similar shape to that obtained by plotting total funding, indicating a large majority of funding comes from the venture capital component.



2.3 Feature Engineering

The `market` and `category_list` broadly cover the same information, `market` is choosen over the information contained in `category_list` due to needing to split/spread the individual categories, using a method such as one hot encoding (adding an extra 824 features).

For the date based variables all three are included, as the plots show periods where the minority statuses of acquired and closed are more significant then usual.

From location based features, `country_code` and `region` are selected. The proportions are very different for companies based in the US, compared to the rest of the world. At the regional level, there are also large differences where the SF Bay Area (Sillicon Valley) has the highest proportion of acquired startups.

Most of the funding based features had zero variation and of no use, the two used are `funding_total_usd` and `funding_rounds`.

One of the models choosen is multinomial logistic classification using the `multinom` function from the `nnet` package. This has the requirement that all features be numeric and in the range [0,1]. So all categorial features are converted to numeric, log10 is taken of `funding_total_usd` due to its large range, and then all values are scaled to be in the range [0,1].

The data is then split into 3 sets, a training set for building the models (70%), a validation set that is used to help the building and evaluation of models (15%), and a test set that is only used to test final model accuracy (15%). Note, this definition of test and validation sets differs from the naming used in the MovieLens project, but is the common usage.

2.4 Data Imbalance

The data is heavily imbalanced with a large prevalence of companies in the operating state. This will make it very difficult to be able to correctly detect the other statuses. To help balance the data, 3

methods will be trialled.

- Undersampling, randomly discarding data that are part of the operating category.
- Oversampling, randomly copying rows from the minority categories to increase their numbers.
- SMOTE, Synthetic Minority Oversampling TEchnique where minority categories have synthetic data created based on existing minority data.

These methods are implemented in the `UBL` package.

2.5 Models

Two models are choosen, multinomial logistic classification, and decision trees.

2.5.1 Multinomial Logistic Classification

Status is considered nominal, not ordinal. Although being closed could be considered the lowest value, it is not obvious whether operating or being acquired is the middle value or the best outcome.

The function `multinom` will be used to fit multinomial log-linear models using neural networks.

Models were created and trained on 4 training sets: original unbalanced data, SMOTE, under sampled, and over sampled. Then predictions were carried out against the validation set.

The confusion matrix for the unbalanced data trained model:

```
Confusion Matrix and Statistics

           Reference
Prediction  operating acquired closed
  operating      3882      332    231
  acquired         53       45     13
  closed           11        1      0

Overall Statistics

               Accuracy : 0.8597
                 95% CI : (0.8493, 0.8696)
    No Information Rate : 0.8638
    P-Value [Acc > NIR] : 0.8003

                  Kappa : 0.1078

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: operating Class: acquired Class: closed
Sensitivity                   0.98378        0.119048      0.000000
Specificity                   0.09486        0.984248      0.997225
Pos Pred Value                0.87334        0.405405      0.000000
Neg Pred Value                0.47967        0.925286      0.946444
Prevalence                    0.86384        0.082750      0.053415
Detection Rate                0.84982        0.009851      0.000000
Detection Prevalence          0.97307        0.024299      0.002627
Balanced Accuracy             0.53932        0.551648      0.498612
```

It has an accuracy close to the no information rate, and is good at predicting the operating status, very poor for the acquired status, and can't detect the closed status at all. The balanced accuracy, taking into account sensitivity and specificity is only slightly better than 0.5 for the first 2 cases, and just around 0.5 for the closed case.

The SMOTE data set model has the following confusion matrix:

```
Confusion Matrix and Statistics

           Reference
Prediction  operating acquired closed
  operating      2734       87     40
  acquired        537      238     59
  closed          675       53    145

Overall Statistics

               Accuracy : 0.6824
                 95% CI : (0.6686, 0.6958)
    No Information Rate : 0.8638
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2675

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: operating Class: acquired Class: closed
Sensitivity                    0.6929         0.62963       0.59426
Specificity                    0.7958         0.85776       0.83164
Pos Pred Value                 0.9556         0.28537       0.16609
Neg Pred Value                 0.2900         0.96251       0.97321
Prevalence                     0.8638         0.08275       0.05342
Detection Rate                 0.5985         0.05210       0.03174
Detection Prevalence           0.6263         0.18257       0.19111
Balanced Accuracy              0.7443         0.74369       0.71295
```

The overall accuracy is much worse here at about 0.68, but does a much better job at predicting the other two statuses. The other models trained on over and under sampled data performed similar to the SMOTE data trained model.

2.5.2 Classification Tree

From the data exploration, it could be seen that there were definite differences in the proportion of status according to market, region etc. It seems like a decision based tree approach would do well at modelling this.

Models were built against the original unbalanced data, and the SMOTE data. The unbalanced data model:

```
Confusion Matrix and Statistics

           Reference
Prediction  operating acquired closed
  operating      3946      378    244
  acquired          0        0      0
  closed            0        0      0
```

```
Overall Statistics

               Accuracy : 0.8638
                 95% CI : (0.8535, 0.8737)
    No Information Rate : 0.8638
    P-Value [Acc > NIR] : 0.5107

                  Kappa : 0

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: operating Class: acquired Class: closed
Sensitivity                    1.0000         0.00000       0.00000
Specificity                    0.0000         1.00000       1.00000
Pos Pred Value                 0.8638             NaN           NaN
Neg Pred Value                    NaN         0.91725       0.94658
Prevalence                     0.8638         0.08275       0.05342
Detection Rate                 0.8638         0.00000       0.00000
Detection Prevalence           1.0000         0.00000       0.00000
Balanced Accuracy              0.5000         0.50000       0.50000
```

As can be seen from the confusion matrix, the model is always predicting the status of operating. For the SMOTE data trained model:

```
Confusion Matrix and Statistics

           Reference
Prediction  operating acquired closed
  operating      2956      124     79
  acquired        491      186     62
  closed          499       68    103

Overall Statistics

               Accuracy : 0.7104
                 95% CI : (0.697, 0.7235)
    No Information Rate : 0.8638
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2406

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: operating Class: acquired Class: closed
Sensitivity                    0.7491         0.49206       0.42213
Specificity                    0.6736         0.86802       0.86887
Pos Pred Value                 0.9357         0.25169       0.15373
Neg Pred Value                 0.2974         0.94986       0.96383
Prevalence                     0.8638         0.08275       0.05342
Detection Rate                 0.6471         0.04072       0.02255
Detection Prevalence           0.6915         0.16178       0.14667
Balanced Accuracy              0.7114         0.68004       0.64550
```

Overall accuracy is slightly better than the multinomial logistic classifier, but it doesn't do as well on predicting the minority classes.

3 Results

The final model chosen is the multinomial logistic classifier trained on balanced (SMOTE) data, it has slightly better accuracy on the minority classes at the cost of lower overall accuracy and prediction of the status of operating. Predicting against the hold out test set gives the following results:

```
Confusion Matrix and Statistics

           Reference
Prediction  operating acquired closed
  operating      2745       90     31
  acquired        561      222     59
  closed          640       66    155

Overall Statistics

               Accuracy : 0.6833
                 95% CI : (0.6696, 0.6968)
    No Information Rate : 0.8636
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2684

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: operating Class: acquired Class: closed
Sensitivity                    0.6956         0.58730       0.63265
Specificity                    0.8058         0.85206       0.83673
Pos Pred Value                 0.9578         0.26366       0.18002
Neg Pred Value                 0.2948         0.95814       0.97573
Prevalence                     0.8636         0.08273       0.05362
Detection Rate                 0.6008         0.04859       0.03392
Detection Prevalence           0.6273         0.18429       0.18844
Balanced Accuracy              0.7507         0.71968       0.73469
```

For comparison, the multinomial model trained on the original unbalanced data has the following results against the test set.

```
Confusion Matrix and Statistics

           Reference
Prediction  operating acquired closed
  operating      3890      340    239
  acquired         45       37      4
  closed           11        1      2

Overall Statistics

               Accuracy : 0.8599
                 95% CI : (0.8495, 0.8699)
    No Information Rate : 0.8636
```

```
    P-Value [Acc > NIR] : 0.7754

                  Kappa : 0.0877

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: operating Class: acquired Class: closed
Sensitivity                   0.98581        0.097884     0.0081633
Specificity                   0.07063        0.988308     0.9972248
Pos Pred Value                0.87044        0.430233     0.1428571
Neg Pred Value                0.44000        0.923935     0.9466520
Prevalence                    0.86365        0.082731     0.0536222
Detection Rate                0.85139        0.008098     0.0004377
Detection Prevalence          0.97811        0.018822     0.0030641
Balanced Accuracy             0.52822        0.543096     0.5026940
```

Essentially there is a trade off in overall accuracy, and accuracy on predicting the majority status, versus accuracy on the minority statuses. The balanced accuracy is much better on the SMOTE trained model. In this project, the use of a rebalancing method was essential to allow the model to be able to predict the minority statuses that had small prevalences of around 8% and 5%.

4 Conclusion

In this project, a multinomial logistic classification model was created to predict the status of startups, based on 8 features such as its market, country, region, funding in USD, founding year etc. The data was heavily unbalanced, with the prevalence of the majority status approximately 86%. Through the use of the SMOTE balancing technique applied to the training data, the model was able to reach a balanced accuracy score of around 0.7 for each of the statuses, but at a cost of lower overall accuracy of approximately 0.68.

The original data was fairly dirty, missing a lot of values and having bad/inconsistent and duplicated rows. There were originally 39 features, however the majority of the funding features were essentially empty giving no predictive power.

Future work could involve evaluating other tree models such as random forests, or gradient boosted decision trees (XGBoost), or applying methods to allow tuning of accuracy for one status class in preference to the others.

Bibliography

Irizarry, Rafael A. 2020. *Introduction to Data Science*. Available at https://rafalab.github.io/dsbook/ (2020/06/23).

Laufer, Jens. n.d. "Missing Value Visualization with Tidyverse in R." Available at https://jenslaufer.com/data/analysis/visualize_missing_values_with_ggplot.html (2020/06/16).

"StartUp Investments (Crunchbase)." 2020. Available at https://www.kaggle.com/arindam235/startup-investments-crunchbase/ (2020/06/16).