Job Salary Prediction

Akshay Gupta	2009EE50275
Prabhav Agrawal	2009EE10405
Akshay Kumar	2008EE50411

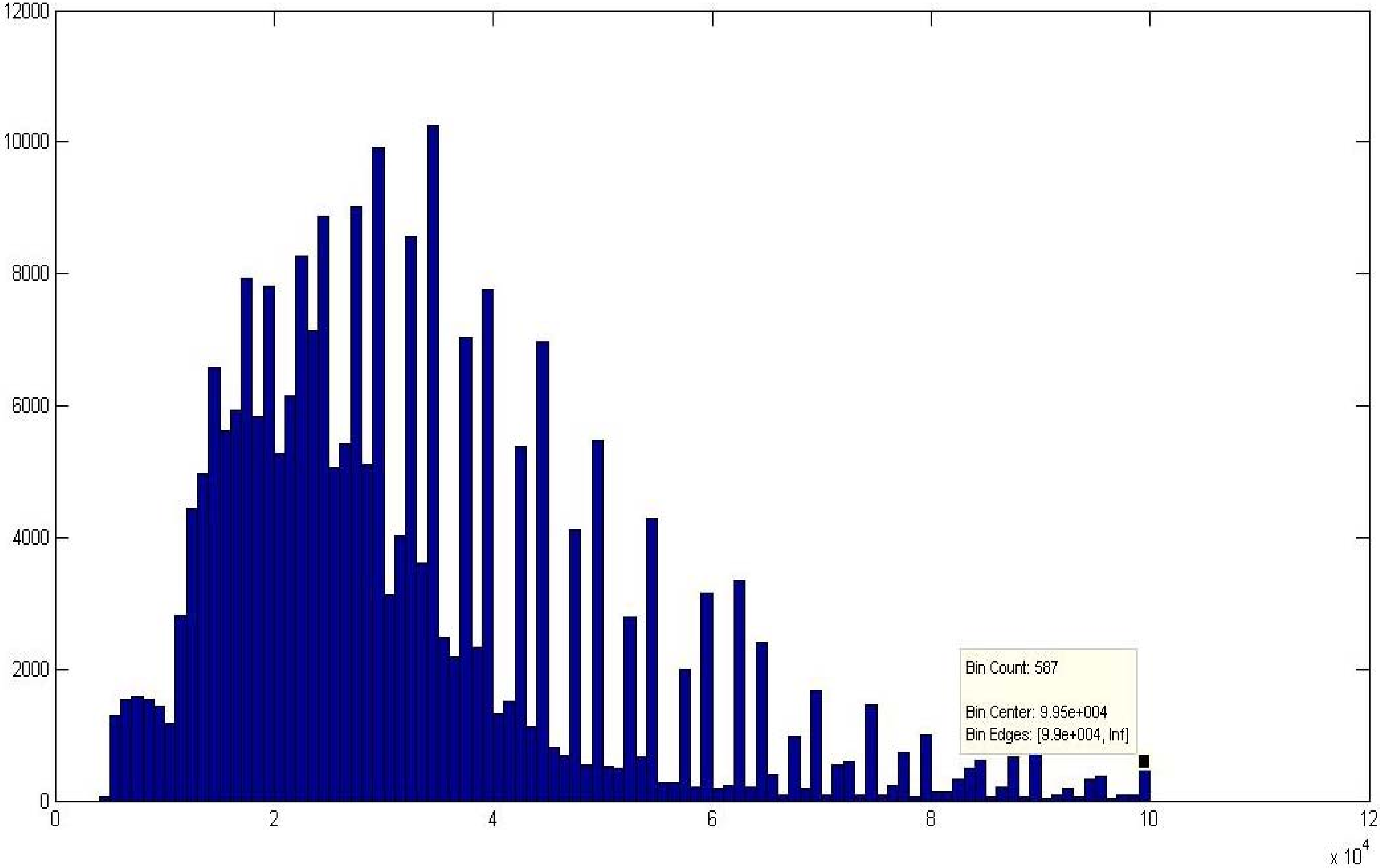
The Data

- Problem: To train a salary prediction engine for UK job advertisements

- The description of each job is given by several fields such as –

1. Title
2. Full Description
3. Company
4. Category
5. Contract Type
6. Location
7. Salary

Histogram of Salaries

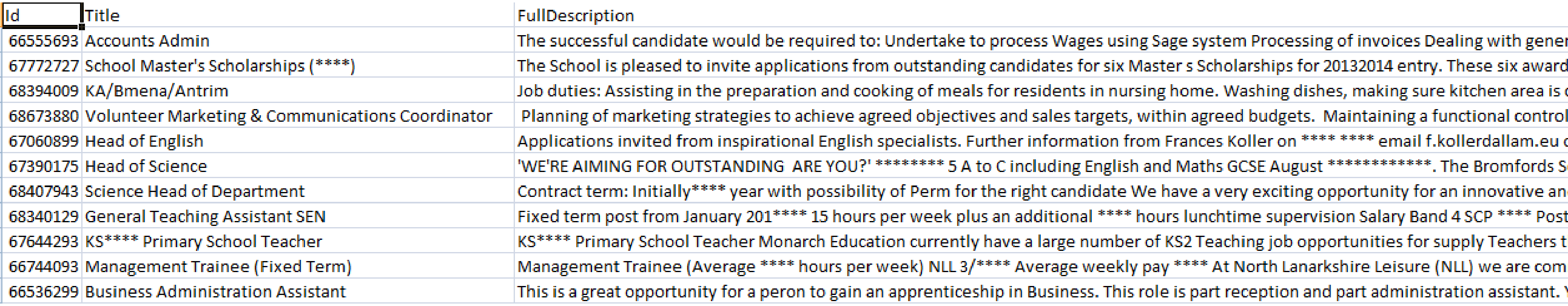


Classifiers & Regressor

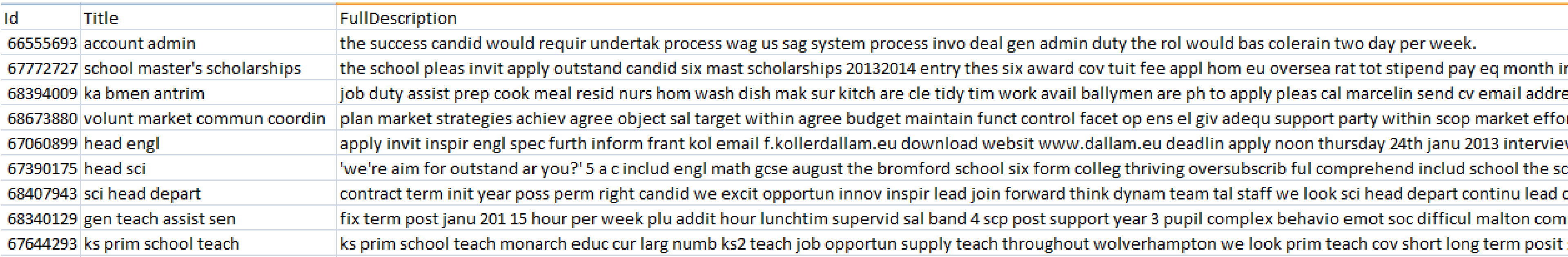
- Number of Classes = 3 (Low, Medium & High)
- Median 1 = 23100
- Median 2 = 34800
- Classifiers : Naïve Bayes & SVM
- Regressor : Random Forest Regressor

Cleaning of Data

- Stemming (NLTk)
- Non ASCII Characters
- Stop Words : English
- Size of Train = 50K Words
- Size of Test = 30k Words



Id	Title	FullDescription
66555693	Accounts Admin	The successful candidate would be required to: Undertake to process Wages using Sage system Processing of invoices Dealing with gener
67772727	School Master's Scholarships (****)	The School is pleased to invite applications from outstanding candidates for six Master s Scholarships for 20132014 entry. These six award
68394009	KA/Bmena/Antrim	Job duties: Assisting in the preparation and cooking of meals for residents in nursing home. Washing dishes, making sure kitchen area is
68673880	Volunteer Marketing & Communications Coordinator	Planning of marketing strategies to achieve agreed objectives and sales targets, within agreed budgets. Maintaining a functional control
67060899	Head of English	Applications invited from inspirational English specialists. Further information from Frances Koller on **** **** email f.kollerdallam.eu
67390175	Head of Science	'WE'RE AIMING FOR OUTSTANDING ARE YOU?' ******** 5 A to C including English and Maths GCSE August ************. The Bromfords S
68407943	Science Head of Department	Contract term: Initially**** year with possibility of Perm for the right candidate We have a very exciting opportunity for an innovative an
68340129	General Teaching Assistant SEN	Fixed term post from January 201**** 15 hours per week plus an additional **** hours lunchtime supervision Salary Band 4 SCP **** Post
67644293	KS**** Primary School Teacher	KS**** Primary School Teacher Monarch Education currently have a large number of KS2 Teaching job opportunities for supply Teachers t
66744093	Management Trainee (Fixed Term)	Management Trainee (Average **** hours per week) NLL 3/**** Average weekly pay **** At North Lanarkshire Leisure (NLL) we are com
66536299	Business Administration Assistant	This is a great opportunity for a peron to gain an apprenticeship in Business. This role is part reception and part administration assistant.



Id	Title	FullDescription
66555693	account admin	the success candid would requir undertak process wag us sag system process invo deal gen admin duty the rol would bas colerain two day per week.
67772727	school master's scholarships	the school pleas invit apply outstand candid six mast scholarships 20132014 entry thes six award cov tuit fee appl hom eu oversea rat tot stipend pay eq month i
68394009	ka bmen antrim	job duty assist prep cook meal resid nurs hom wash dish mak sur kitch are cle tidy tim work avail ballymen are ph to apply pleas cal marcelin send cv email addre
68673880	volunt market commun coordin	plan market strategies achiev agree object sal target within agree budget maintain funct control facet op ens el giv adequ support party within scop market effo
67060899	head engl	apply invit inspir engl spec furth inform frant kol email f.kollerdallam.eu download websit www.dallam.eu deadlin apply noon thursday 24th janu 2013 intervie
67390175	head sci	'we're aim for outstand ar you?' 5 a c includ engl math gcse august the bromford school six form colleg thriving oversubscrib ful comprehend includ school the sc
68407943	sci head depart	contract term init year poss perm right candid we excit opportun innov inspir lead join forward think dynam team tal staff we look sci head depart continu lead
68340129	gen teach assist sen	fix term post janu 201 15 hour per week plu addit hour lunchtim supervid sal band 4 scp post support year 3 pupil complex behavio emot soc difficul malton com
67644293	ks prim school teach	ks prim school teach monarch educ cur larg numb ks2 teach job opportun supply teach throughout wolverhampton we look prim teach cov short long term posit

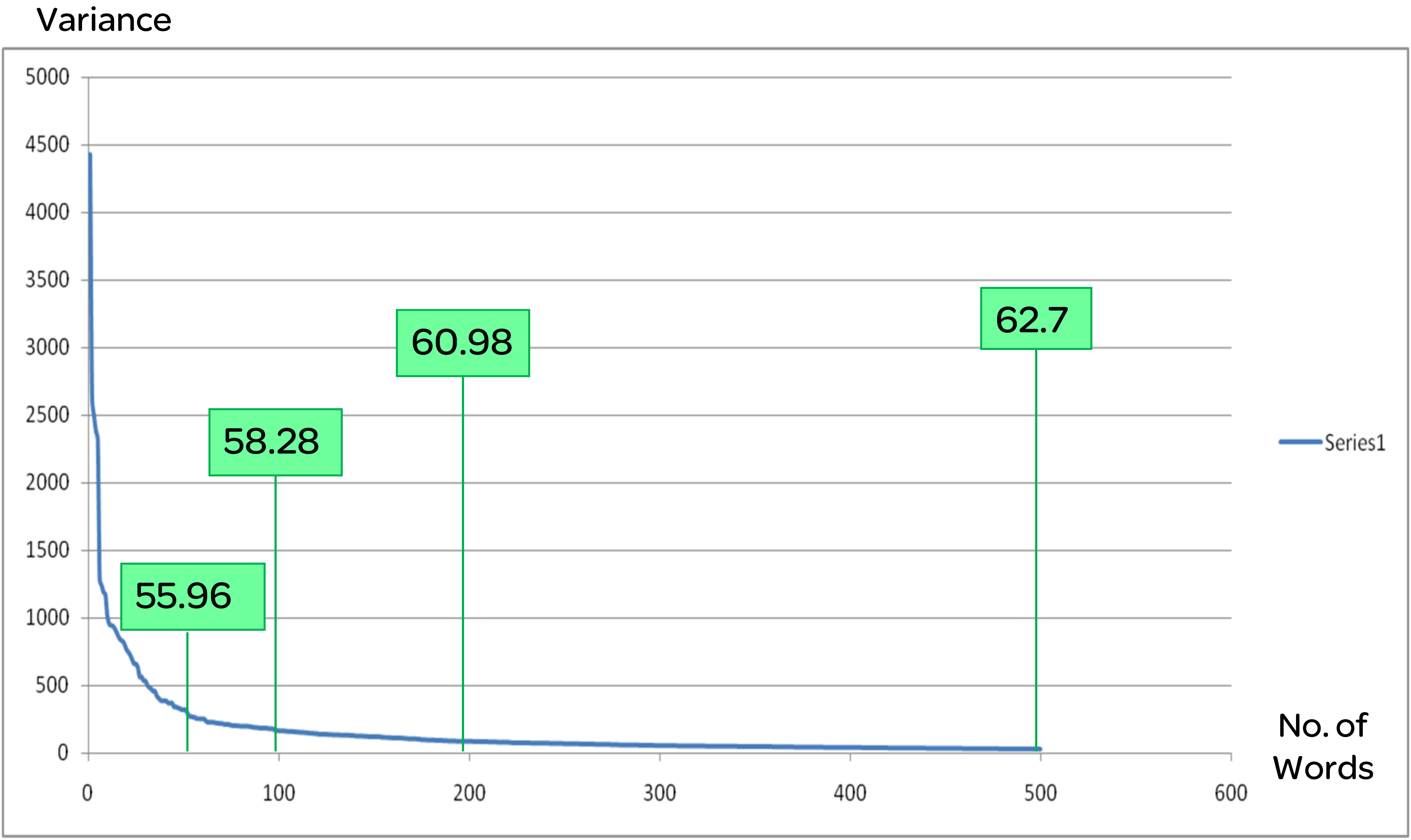
Title & Full Description

- Calculated Variance of each word and then applied bag of words with dictionary containing words with high variance

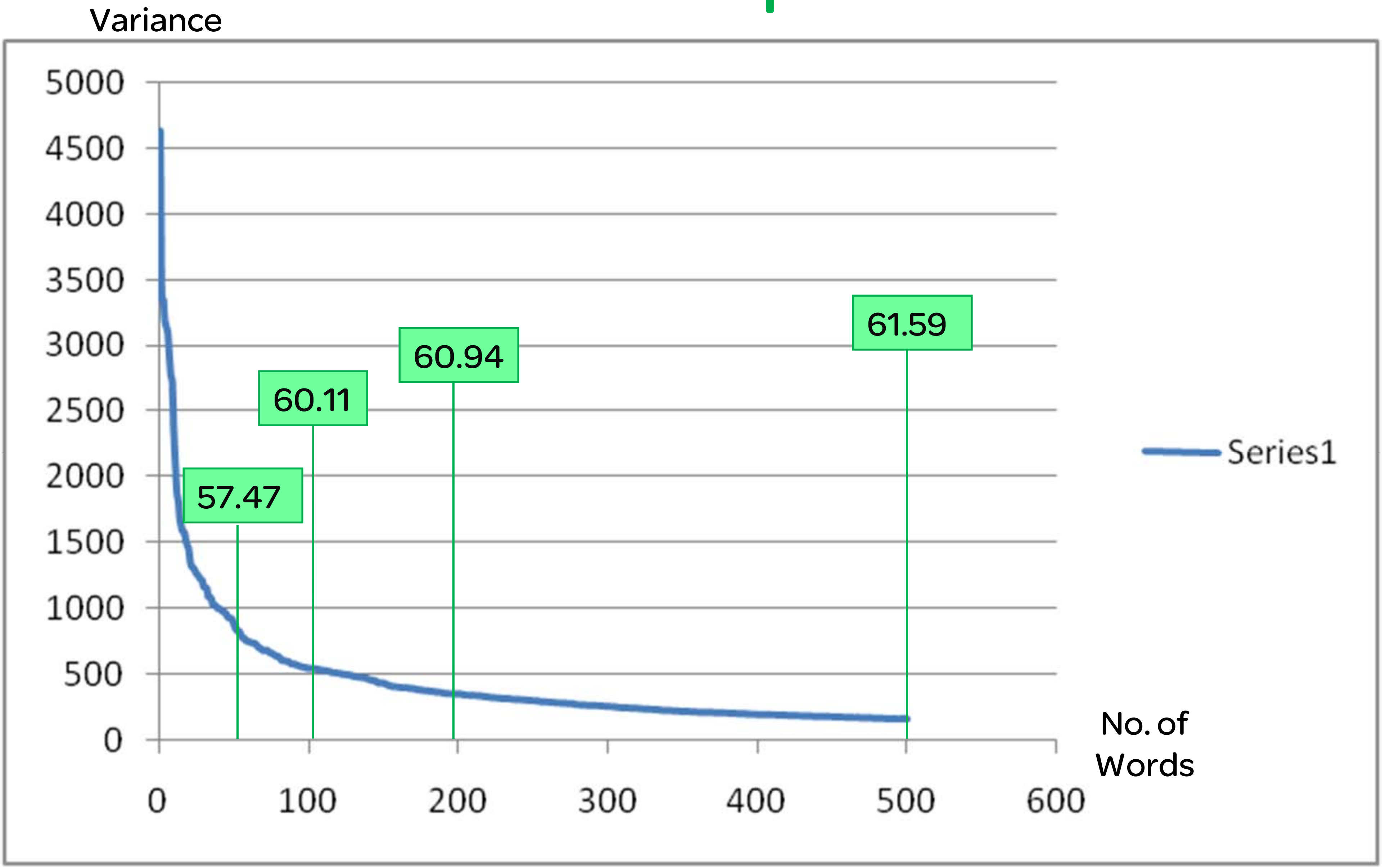
Title
Man
Assist
Engin
Develop
Admin

Full Description
Work
Develop
Project
Engin
Man

Title


Variance
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
No. of Words
Series1
55.96
58.28
60.98
62.7

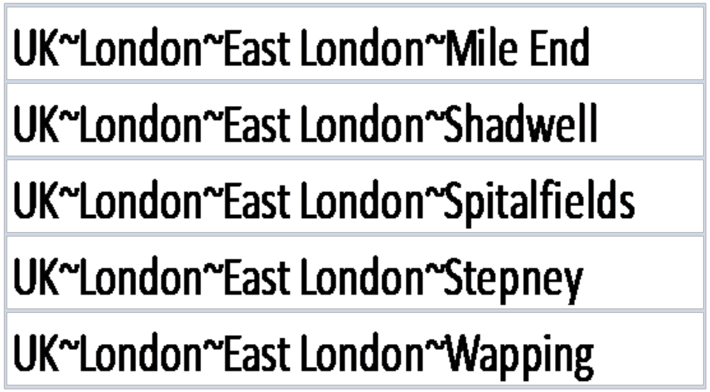
Full Description



Locations

- Given in the form of a location tree
- **Level 1** – 18 Level1 Locations; Locations given in the data are not unique and similar locations have been merged
- **Level 2**- 241 Locations
- **Total**- 31763 Locations

UK~Level1~Level2~Level3



UK~London~East London~Mile End
UK~London~East London~Shadwell
UK~London~East London~Spitalfields
UK~London~East London~Stepney
UK~London~East London~Wapping

Frequency	Place	Avg. Salary	Phigh	Pmedium	Plow
11822	'london'	41330	0.56685	0.282875	0.150275
1658	'scotland'	30990	0.308248	0.295003	0.396749
844	'north ireland'	29140	0.293979	0.353011	0.353011
3608	'south west england'	28400	0.263085	0.307671	0.429244
2740	'east england'	31650	0.356544	0.278892	0.364564
8585	'south east england'	31050	0.333489	0.300303	0.366209
574	'wales'	27010	0.244367	0.268631	0.487002
3375	'north west england'	27310	0.234754	0.287152	0.478094
2735	'east midlands'	28180	0.252739	0.327977	0.419284
1444	'north east england'	27400	0.234969	0.296475	0.468556
1847	'yorkshire and humberside'	28490	0.255135	0.307568	0.437297
2941	'west midlands'	30000	0.302989	0.332201	0.36481
5	'isle of man'	45240	0.5	0.375	0.125
11	'channel islands'	37240	0.428571	0.285714	0.285714

Mean	32700
Median	28800
Median1	23100
Median 2	34800

Company Names

- 8998 distinct company names.

Company	Frequency	Avg. Salary	Company	Frequency	Avg. Salary
'ukstaffsearch'	1192	34220	'support services group'	193	29380
'cvbrowser'	974	31730	'chef results'	181	18700
'london4jobs'	496	40340	'perfect placement'	179	25560
'hays'	485	29410	'corecruitment international'	174	32160
'jobg8'	466	41470	'computer people'	173	42110
'array'	445	30180	'clear selection'	168	20470
'fresh partnership'	297	27780	'cmc consulting'	168	50210
'matchtech group plc'	292	44650	'triumph consultants'	160	30510
'jam recruitment'	231	45550	'randstad'	158	33300
'office angels'	203	20490	'sf group'	145	29900

Company Names Words

- 7069 distinct words occur in company names

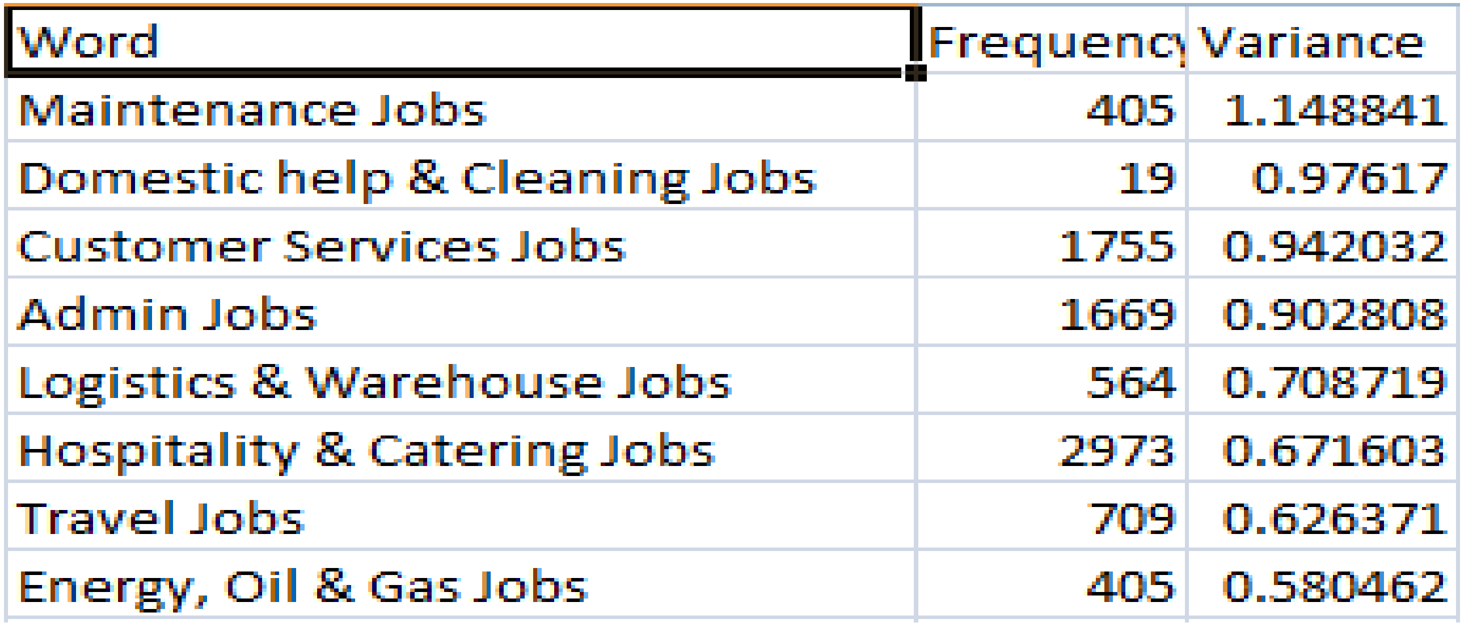
Word	Frequency	Avg. Salary	Word	Frequency	Avg. Salary
'recruitment'	10469	31650	'cvbrowser'	974	31730
'limited'	9767	33360	'&'	924	33190
'group'	2716	33020	'resourcing'	911	36110
'solutions'	1494	33550	'selection'	786	28990
'services'	1403	27970	'people'	784	35840
'ukstaffsearch'	1192	34220	'care'	720	23740
'personnel'	1088	27760	'it'	704	42720
'hays'	1064	34740	'uk'	672	29510
'associates'	1032	37020	'plc'	663	37350
'consulting'	1027	43650	'international'	625	36710

Company Names Words

Word	Avg. Salary	Word	Avg. Salary	Word	Avg. Salary	Word	Avg. Salary
'moran'	55210	'masson'	53110	'austin'	47260	'badenoch'	45610
'ruth'	54670	'real'	51000	'global'	46270	'alexander'	45200
'partners'	54300	'cmc'	50210	'morgan'	46160	'jam'	45080
'experis'	53140	'technology'	48110	'modis'	45920	'lloyd'	45050
'goodman'	53110	'senior'	48070	'harvey'	45840	'project'	44630
'cleaning'	13030	'office'	20100	'bupa'	21920	'trust'	22800
'dot'	17670	'trade'	20210	'school'	22230	'careers'	22830
'results'	18720	'angels'	20490	'acs'	22370	'four'	22840
'home'	18900	'adecco'	21010	'seasons'	22470	'travel'	23180
'chef'	19010	'towngate'	21770	'staff'	22660	'interaction'	23180

Location & Category

- These both are categorical features:
  - Location: 18 (Level 1)
  - Category: 28 Categories



Word	Frequency	Variance
Maintenance Jobs	405	1.148841
Domestic help & Cleaning Jobs	19	0.97617
Customer Services Jobs	1755	0.942032
Admin Jobs	1669	0.902808
Logistics & Warehouse Jobs	564	0.708719
Hospitality & Catering Jobs	2973	0.671603
Travel Jobs	709	0.626371
Energy, Oil & Gas Jobs	405	0.580462

Feature	Accuracy
Location	39
Category	46.63

Combination of Features

Combination	Accuracy
Title + Full Description	62.68
Title + Category	61.45
Title + Location	61.97
Full Description + Location	60.32
Full Description + Category	60.44
Category + Location	49.37

Results

- Title : 200 words
- Full Description : 100 words
- Locations : 18
- Category : 28

Classifier/ Regressor	Accuracy/ R^2
Naïve Bayes	63.03
SVM	68.07
Random Forest (Regressor)	0.49

Thank You!!!