Classification of bank-accout transactions & bank-account balance time-series prediction

Date: 7.12.2014
Author: Niklas Kolster, email: niklas.kolster@gmail.com

Classification of bank-account transactions into categories

Background

We have two csv files, one consisting of categories and one consisting of bank-account transactions. Ca. 50% of the transactions have been classified into one of the sixteen categories. Our task is to look into classifying the rest of the transactions.

We proceed with loading and looking into the data:

```
# Loading some libraries
library(xts)
library(quantmod)
# Going to the directory with the files
setwd("~/Folder_with_data/") # change path accordingly
# Reading in the categories
cats <- read.table("categories.csv", header=T, sep=",")
```

The categories look like this:

id	name
10	Bargeld
11	Elektronik
12	Gesundheit
13	Handy & Internet
14	Haushalt
15	Kinder
16	Kleidung
17	Lebensmittel
18	Medien
19	Miete & Hypothek
20	Reisen & Urlaub
21	Restaurants & Freizeit
22	Sport & Hobby
23	Transport
24	Versicherungen
25	Wohnnebenkosten
26	Sonstiges
27	Nicht kategorisiert

Reading in the transactions

```
tx <- read.table("anonymized-test-transactions.csv", header=T,
sep=",",stringsAsFactors=F)
```

A category_id below 27 means the data has been classified and can be used as training # data.

```
tr <-tx[tx$category_id < 27,]
```

Category_id 27 means the data has not been classified so it is used as test data, this is our "new data"

```
test <- tx[tx$category_id == 27,]
```

Now the data has been split up in training (tr) and test (test) data.
We want to get a quick look at how the data is composed.

The data looks like this:

```
> head(tx)
        id         booking_date      effective_date gvcode                                                                                   usage additional transaction_type user_bank_account_id category_id
1 30456709 2014-03-17 00:00:00 2014-03-17 00:00:00   NULL                                      EC 78101424 160314184153IC6 A3E154E763A292AA409E2564D1       NULL     KARTENVERF?G              7350408          23
2 30456710 2014-03-17 00:00:00 2014-03-17 00:00:00   NULL                                       QVZJ7Z/DE08600100700667567 01 RG 013-14 Casper Tour       NULL      ?BERWEISUNG              7350408          27
3 30456711 2014-03-14 00:00:00 2014-03-14 00:00:00   NULL                                                                696909168016 WATCHEVER GMBH       NULL      LASTSCHRIFT              7350408          26
4 30456712 2014-03-13 00:00:00 2014-03-13 00:00:00   NULL                                          Referenz NOTPROVIDED Verwendungszweck RG 013-14       NULL       GUTSCHRIFT              7350408          27
5 30456713 2014-03-12 00:00:00 2014-03-12 00:00:00   NULL KBS 83132249 KRT0006/12.18 12.03 14.01 TA-NR. 146840 99084 Erfurt Anger 66-73 EC-CARD MIT PIN       NULL             NULL              7350408          27
6 30456714 2014-03-12 00:00:00 2014-03-12 00:00:00   NULL                                                 NYEF7Q/DE08600100700667567 01 sparen       NULL      ?BERWEISUNG              7350408          27
  contact_bank_account_id transaction_group_id                          id_hash balance   amount contact_id
1                 7349618                 NULL 0ced49f91885fe61762b7db5eefbdb11  8857.3   -28.93    7666703
2                    NULL                 NULL 51bfa9a1488e2c504736f274581acbf9  8886.2 -1530.83    7666650
3                 7349621                 NULL b5578e456a087ab94e702b513c5b88ca 10417.0    -8.99    7666706
4                 7349571                 NULL 4504dbce51358ccba87f2882e775cdee 10426.0  9857.84    7666655
5                    NULL                 NULL a654495c22b4215d67c51b9097f8ee86   568.2 -4050.00    7667509
6                    NULL                 NULL 59c4122880692d9b820dc4931eba1baa  4618.2 -4000.00    7666650
```

Full size picture: http://i.imgur.com/mMxkSqm.png

The data has the following columns:

```
colnames(tx)
"id" "booking_date" "effective_date" "gvcode" "usage" "additional"
"transaction_type" "user_bank_account_id" "category_id"
"contact_bank_account_id" "transaction_group_id" "id_hash" "balance" "amount"
"contact_id"
```

Rows in training:

```
nrow(tr)
14892
```

Rows in test:

```
nrow(test)
21454
```

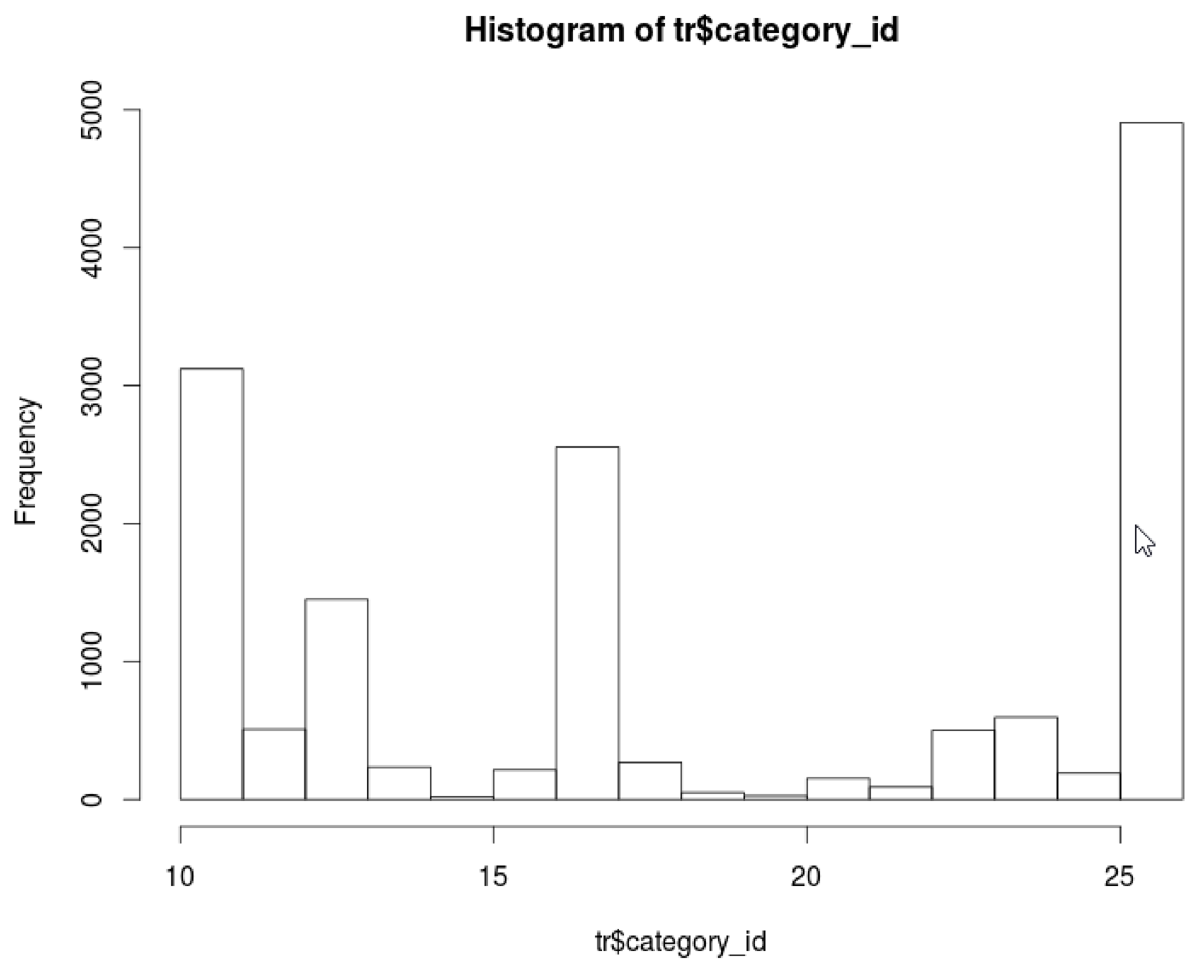
Split:

```
nrow(tr)/nrow(tx)
0.4097287
```

So it seems ca. 40% of the data is classified

Getting initial understanding of the distributions between categories

```
hist(tr$category_id):
```



Looking at the distribution numbers from category perspective

```
sort(table(tr$category_id), decreasing=T)
26   10   17   13   24   12   23   18   14   11   16   25   21   22   19   20   15
4905 2907 2555 1451  596  509  503  269  235  215  215  191  153   91   50   27   20
```

Some of the categories are large so we will focus first on them where we have the most data to work with.

Looking at the contact_id's

```
head(sort(table(tr$contact_id), decreasing=T))
7670951 7669873 7671704 7679644 7669881 7670534
270     159     145     140     119     116
```

We proceed to take look at the largest groups of contact_id’s below.

```
# View(tr[tr$contact_id==7670951,])
```

These lines stand out:

usage	additional	transaction_type	user_bank_account_id	category_id	contact_bank_account_id	transaction_group_id	id_hash	balance	amount
MASTERCARD 5232540875138285 PAYPAL *MICROSOFTLU 35314369001 24.04.2014 9.99 EUR	NULL	EIGENE KREDITKARTENABRECHN.	7353673	18	7353678	NULL	ba0bd06cc536246a8c6260d63d2ee9ed	523.7	-9.99
MASTERCARD 5232540875138285 APPLE ITUNES STORE-EUR ITUNES.COM 26.04.2014 1.78 EUR	NULL	EIGENE KREDITKARTENABRECHN.	7353673	18	7353678	NULL	eceebb7b0be8f37c12b4920b350dad1b	533.7	-1.78
MASTERCARD 5232540875138285 PAYPAL *MICROSOFTLU 35314369001 16.04.2014 3.49 EUR	NULL	EIGENE KREDITKARTENABRECHN.	7353673	26	7353678	NULL	ebef3de6c588b56128423c1c9c795450	293.5	-3.49
MASTERCARD 5232540875138285 APPLE ITUNES STORE-EUR ITUNES.COM 17.04.2014 6.48 EUR	NULL	EIGENE KREDITKARTENABRECHN.	7353673	26	7353678	NULL	a3919db8460931441070df9328d7f583	297.0	-6.48

Link to sharper picture: http://i.imgur.com/XXJZB26.png

Some i-tunes are in media, some in sonstiges etc.. **Why?**
This raises some questions about the quality of the training data.

Looking at different contact_id's below to get a feel for them, with some comments.
Pictures not shown here to save space.

```
View(tr[tr$contact_id==7669873,])
#All in sonstiges

View(tr[tr$contact_id==7671704,])
# All in sonstiges
# User bank account varies for this. Is that possible for the same contact ID?
```

Looking at different categories to see of something clearly stands out:

```
View(tr[tr$category_id==26,])
# This (Sonstiges=Misc) is not a very intresting category although the largest.
Also has itunes for some reason.

View(tr[tr$category_id==17,])
# Lebensmittel is large but at first look no clear tell-tales for easy rules

View(tr[tr$category_id==13,])
# Contains EREF etc. Mabye contact id can help here, could check.

View(tr[tr$category_id==10,])
# Bargeld are even numbers and contain UHR, type is usually geldautomat or
auszahlung
```

The view of category 10 (Bargeld=Cash):

row.names	id	booking date	effective date	gvcode	usage	additional	transaction type	user bank account id	category id	contact bank account id	transaction group id	id hash	balance	amount	contact id
112	30456829	2014-05-02 00:00:00	2014-05-02 00:00:00	NULL	Verwendungszweck sparen	NULL	GUTSCHRIFT	7350409	10	7350414	NULL	baf069459960e9e520ba5b5d5fad5da9	21416.25	7500.0	7667511
113	30456830	2014-05-02 00:00:00	2014-05-02 00:00:00	NULL	Verwendungszweck Casper RG 15/16-14	NULL	GUTSCHRIFT	7350409	10	7350414	NULL	f787b76c0b4a992c154b5297176ad668	13916.25	1916.5	7667511
234	30457276	2014-05-07 00:00:00	2014-05-07 00:00:00	NULL	07.05/13.49UHR ARNERFSTR	NULL	GELDAUTOMAT	3632554	10	3632559	NULL	276b8bca69718ad00fefcea9f15646d3	-81.00	-20.0	3799855
260	30457309	2014-03-03 00:00:00	2014-03-01 00:00:00	NULL	01.03/22.45UHR KOEBRUESTR	NULL	GELDAUTOMAT	7350593	10	7350613	NULL	bcbcd107a98101cf6dc77fb4626b15ac	120.82	-25.0	7667719
269	30457319	2014-02-04 00:00:00	2014-02-04 00:00:00	NULL	04.02/17.09UHR WEISSPARK	NULL	GELDAUTOMAT	7350593	10	7350617	NULL	30f962df80830addc942a17600e2a7f1	15.23	-10.0	7667723
288	30457345	2013-12-09 00:00:00	2013-12-07 00:00:00	NULL	07.12/22.42UHR SBALBERTPL	NULL	GELDAUTOMAT	7350593	10	7350625	NULL	35c40b2236aac39ded8b2c03c2233d1b	122.63	-40.0	7667731
296	30457355	2013-11-15 00:00:00	2013-11-15 00:00:00	NULL	15.11/16.19UHR THARANDT	NULL	GELDAUTOMAT	7350593	10	7350629	NULL	6137598919b2fecd39ec1a1e5aede504	28.58	-10.0	7667735
319	30457390	2013-08-01 00:00:00	2013-08-01 00:00:00	NULL	01.08/12.11UHR FTLPOTSCH	NULL	GELDAUTOMAT	7350593	10	7350637	NULL	3f07cf05ff3ae11cfc4a2dc47cc4b88d	9.81	-25.0	7667743
334	30457410	2013-06-05 00:00:00	2013-06-05 00:00:00	NULL	05.06/13.43UHR THARANDT	NULL	GELDAUTOMAT	7350593	10	7350629	NULL	349e8cc82033cc5b10ba86a25e4db07e	16.81	-10.0	7667735
372	30457459	2012-12-27 00:00:00	2012-12-27 00:00:00	NULL	27.12/19.15UHR WEISSPARK	NULL	GELDAUTOMAT	7350593	10	7350648	NULL	676e115c6db53ca5a2662a57db5c6bfd	9.20	-90.0	7667754
387	30457476	2012-11-30 00:00:00	2012-11-29 00:00:00	NULL	29.11/20.03UHR FTLPOTSCH	NULL	GELDAUTOMAT	7350593	10	7350637	NULL	d228cc86a65ac959093c9030ff6bd3fe	76.96	-10.0	7667743
862	30458213	2014-05-02 00:00:00	2014-05-02 00:00:00	NULL	02.05/12:52 UHR Ettishofen	NULL	SB-Auszahlung	7350819	10	7350822	NULL	2bacbb03eeb59c64a5b193f098231f39	668.29	-50.0	7667920
871	30458222	2014-04-22 00:00:00	2014-04-20 00:00:00	NULL	20.04/17:38 UHR Oberhofen	NULL	SB-Auszahlung	7350819	10	7350829	NULL	50ca125e5ab855f58b816336244fed26	415.52	-50.0	7667928
877	30458228	2014-04-09 00:00:00	2014-04-09 00:00:00	NULL	09.04/18.26UHR Gr?nkraut	NULL	SB-Auszahlung	7350819	10	7350833	NULL	b6cfc766b602af9c2359946dc3c9c214	614.03	-250.0	7667932
881	30458232	2014-04-02 00:00:00	2014-04-02 00:00:00	NULL	02.04/19.28UHR Gr?nkraut	NULL	SB-Auszahlung	7350819	10	7350833	NULL	29428e87ca256ffd8dac2475c84c905f	904.39	-20.0	7667932
884	30458236	2014-03-31 00:00:00	2014-03-31 00:00:00	NULL	31.03/12.10UHR RV Bachstr	NULL	SB-Auszahlung	7350819	10	7350836	NULL	f9588658ade31d3541090e2ce118fd3a	879.21	-20.0	7667935
902	30458254	2014-02-27 00:00:00	2014-02-26 00:00:00	NULL	26.02/23.01UHR Gr?nkraut	NULL	SB-Auszahlung	7350819	10	7350845	NULL	ce823bea4b99492bd97036f497b26104	196.95	-100.0	7667944
905	30458257	2014-02-24 00:00:00	2014-02-22 00:00:00	NULL	22.02/17:05 UHR Oberhofen	NULL	SB-Auszahlung	7350819	10	7350829	NULL	cba3cf70fa6e1d2501015fb1f062c8e0	324.18	-100.0	7667928
912	30458265	2014-02-03 00:00:00	2014-02-01 00:00:00	NULL	01.02/17:33 UHR Horgenzell	NULL	SB-Auszahlung	7350819	10	7350849	NULL	761f7acd85829faff411a9508d0ba636	585.78	-50.0	7667948
916	30458271	2014-01-27 00:00:00	2014-01-25 00:00:00	NULL	25.01/10:23 UHR Horgenzell	NULL	SB-Auszahlung	7350819	10	7350850	NULL	689e469ce53e177008210684bac27480	163.02	-50.0	7667949
919	30458274	2014-01-15 00:00:00	2014-01-15 00:00:00	NULL	15.01/13.05UHR RV Bachstr	NULL	SB-Auszahlung	7350819	10	7350836	NULL	bf1a584abab28afb278989c7a0a5664b	253.03	-40.0	7667935
1038	30458416	2014-02-17 00:00:00	2014-02-15 00:00:00	NULL	15.02/14.59UHR FAULBRUNN.	NULL	GELDAUTOMAT	7350864	10	7350900	NULL	ea56d4ce2239a655aa7c3ee7a0b93432	1041.93	-350.0	7668000
1056	30458440	2014-01-20 00:00:00	2014-01-19 00:00:00	NULL	19.01/16.08UHR FAULBRUNN.	NULL	GELDAUTOMAT	7350864	10	7350900	NULL	bf67305b40ad81de92435e112a340218	3178.39	-1500.0	7668000
1072	30458457	2014-05-07 00:00:00	2014-05-07 00:00:00	NULL	07.05/09.46UHR TABAKLADEN	NULL	GELDAUTOMAT	885907	10	885922	NULL	0408fd2908197e3320d4f0290074cf60	97.21	-70.0	942061
1075	30458460	2014-05-07 00:00:00	2014-05-07 00:00:00	NULL	07.05/09.46UHR TABAKLADEN	NULL	GELDAUTOMAT	6041248	10	885922	NULL	0408fd2908197e3320d4f0290074cf60	97.21	-70.0	942061
1096	30458487	2014-05-05 00:00:00	2014-05-04 00:00:00	NULL	04.05/19.13UHR KOW	NULL	GELDAUTOMAT	4375218	10	4375227	NULL	221a8a382a8f71b34313e1add6cd1a9d	868.08	-520.0	4571487
1270	30458678	2014-05-07 00:00:00	2014-05-07 00:00:00	NULL	07.05/09.34UHR 001-CRS-01	NULL	GELDAUTOMAT	6654502	10	6654516	NULL	b676ee0d57582c18af54e414a71b725b	-102.86	-10.0	6944614
1272	30458680	2014-05-06 00:00:00	2014-05-06 00:00:00	NULL	06.05/17.33UHR 001-CRS-01	NULL	GELDAUTOMAT	6654502	10	6654516	NULL	a027a3e97bd38e9d2a85374299050548	-66.94	-80.0	6944614
1644	30459140	2014-05-02 00:00:00	2014-05-01 00:00:00	NULL	01.05/18.33UHR OSTHEIM	NULL	GELDAUTOMAT	7351148	10	7351152	NULL	c3fb9ac69c3686b63b3893cf0402671c	-527.80	-40.0	7668287
1652	30459150	2014-04-23 00:00:00	2014-04-23 00:00:00	NULL	23.04/19.07UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	86aa30d8e7fb93d6e4bc34b0751411cd	-147.66	-200.0	7668293
1655	30459156	2014-04-17 00:00:00	2014-04-17 00:00:00	NULL	17.04/18.04UHR VINGST	NULL	GELDAUTOMAT	7351148	10	7351161	NULL	13241c706b6bd418b1d58fbfbdd864a8	-476.02	-60.0	7668297
1666	30459174	2014-04-14 00:00:00	2014-04-12 00:00:00	NULL	12.04/18.03UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	0c2e0946910fb36a6dee6b459878b051	-456.46	-200.0	7668293
1667	30459175	2014-04-14 00:00:00	2014-04-11 00:00:00	NULL	11.04/21.19UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	bdccca38e9c9c6cb308aa257d8142516	-256.46	-45.0	7668293
1668	30459176	2014-04-11 00:00:00	2014-04-11 00:00:00	NULL	11.04/11.24UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	e1e1390cf7845adc172f86c9b1f4c721	-211.46	-700.0	7668293
1684	30459202	2014-03-17 00:00:00	2014-03-14 00:00:00	NULL	14.03/21.26UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	c449f4cb2fd5c7f2368012c54258fcb9	-291.56	-540.0	7668293
1704	30459234	2014-02-17 00:00:00	2014-02-15 00:00:00	NULL	15.02/17.16UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	4c135a6909b6549ba29f97aa652271dc	-392.08	-100.0	7668293
1705	30459238	2014-02-13 00:00:00	2014-02-13 00:00:00	NULL	13.02/07.51UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	f036793f9addb7f59c774e85406cf704	159.92	-560.0	7668293
1709	30459242	2014-02-11 00:00:00	2014-02-11 00:00:00	NULL	11.02/18.11UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	412d58216b6fffcb4171bb7f5d5f015e	-534.50	-140.0	7668293
1712	30459245	2014-02-06 00:00:00	2014-02-06 00:00:00	NULL	06.02/13.27UHR NEU-BRUECK	NULL	GELDAUTOMAT	7351148	10	7351189	NULL	5f1c86f08741ac16d53202b1b297f436	-596.17	-15.0	7668325
1713	30459246	2014-02-05 00:00:00	2014-02-05 00:00:00	NULL	05.02/19.03UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	9060d35b2d1f77c13edb3623994cce05	-581.17	-90.0	7668293
1754	30459301	2013-11-28 00:00:00	2013-11-28 00:00:00	NULL	28.11/11.09UHR NEU-BRUECK	NULL	GELDAUTOMAT	7351148	10	7351189	NULL	e5e39b069ed9c74e00e213d63de9dc0d	[illegible]	-40.0	7668325
1769	30459324	2013-10-30 00:00:00	2013-10-30 00:00:00	NULL	30.10/13.29UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	17910a679a4f1e44d92d90f67224cbb1	-175.14	-900.0	7668293
1778	30459343	2013-10-16 00:00:00	2013-10-16 00:00:00	NULL	16.10/16.04UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	8c9a67f793b14f2f2089480196a83125	-497.68	-90.0	7668293
1789	30459357	2013-09-23 00:00:00	2013-09-21 00:00:00	NULL	21.09/11.07UHR KALK	NULL	GELDAUTOMAT	7351148	10	7351215	NULL	8b4d0cd5aa792ab2a6d87521ead5505c	-464.55	-5.0	7668351
1798	30459369	2013-08-26 00:00:00	2013-08-24 00:00:00	NULL	24.08/14.13UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351157	NULL	9867a14f883921288fb0ab8806da1fdc	-529.39	-150.0	7668293
1802	30459375	2013-08-05 00:00:00	2013-08-03 00:00:00	NULL	03.08/09.45UHR MERHEIM	NULL	GELDAUTOMAT	7351148	10	7351221	NULL	8a420d5aa0fce85b909c4a1f56c3ab14	-594.76	-15.0	7668357
1813	30459387	2013-07-22 00:00:00	2013-07-21 00:00:00	NULL	21.07/13.58UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351224	NULL	10e0cdc47c67f2e9b26f82f4761b094e	-538.58	-15.0	7668360
1822	30459407	2013-07-02 00:00:00	2013-07-02 00:00:00	NULL	02.07/17.19UHR NEU-BRUECK	NULL	GELDAUTOMAT	7351148	10	7351228	NULL	6166f9c9cce74ce35428f8c759077488	-595.55	-30.0	7668364
1824	30459409	2013-07-01 00:00:00	2013-06-29 00:00:00	NULL	29.06/15.23UHR KALK	NULL	GELDAUTOMAT	7351148	10	7351215	NULL	6a65f570bc6dbfcddda58d3e0f8d5041	-565.55	-30.0	7668351
1845	30459440	2013-05-28 00:00:00	2013-05-28 00:00:00	NULL	28.05/14.00UHR KALK	NULL	GELDAUTOMAT	7351148	10	7351215	NULL	432ff510fca3a7cab3fdc50d6f08401e	-566.38	-100.0	7668351
1849	30459445	2013-05-27 00:00:00	2013-05-26 00:00:00	NULL	26.05/20.46UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351224	NULL	b313a7d3a42a1a31cbc50bca7ce9dd28	-382.58	-100.0	7668360
1851	30459447	2013-05-23 00:00:00	2013-05-22 00:00:00	NULL	22.05/22.16UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351224	NULL	98c9ee222f266e9dfb95553786d31c8a	-580.79	-100.0	7668360
1853	30459450	2013-05-17 00:00:00	2013-05-17 00:00:00	NULL	17.05/19.17UHR RODENKIRCH	NULL	GELDAUTOMAT	7351148	10	7351235	NULL	78790e2ee4b818feac7c9107f731d8cd	-525.79	-30.0	7668371
1858	30459459	2013-05-10 00:00:00	2013-05-10 00:00:00	NULL	10.05/11.37UHR KALK	NULL	GELDAUTOMAT	7351148	10	7351236	NULL	15c259765f360a2aa77c4b97432ff0ae	-594.00	-500.0	7668372
1863	30459467	2013-05-03 00:00:00	2013-05-03 00:00:00	NULL	03.05/11.56UHR NEU-BRUECK	NULL	GELDAUTOMAT	7351148	10	7351228	NULL	e775702e21eb820fc0888ae74705a13f	-495.15	-100.0	7668364
1864	30459468	2013-05-03 00:00:00	2013-05-02 00:00:00	NULL	02.05/21.41UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351224	NULL	6773581468611f73d52c976cd23900f9	-366.18	-210.0	7668360
1865	30459469	2013-05-02 00:00:00	2013-05-02 00:00:00	NULL	02.05/00.05UHR RONDORF	NULL	GELDAUTOMAT	7351148	10	7351238	NULL	a34935fb28609ab12666ddb4c0abaec6	-156.18	-100.0	7668374
1877	30459484	2013-04-02 00:00:00	2013-03-30 00:00:00	NULL	30.03/16.03UHR GODORF	NULL	GELDAUTOMAT	7351148	10	7351224	NULL	d88fe4f85525968af7fdcee463d82ccd	-599.74	-100.0	7668360

Link to full-size: http://i.imgur.com/LBkpici.png

Forming simple rule for Bargeld (Category 10):

Looking into forming straightforward rule for bargeld.
In Training data transaction_type has either geldautomat or auszahlung for most items in Bargeld.

Category 10:

```
tr$geldautomat_or_auszahlung <- ( grepl("geldautomat", tr$transaction_type,
ignore.case=T) | grepl("auszahlung", tr$transaction_type, ignore.case=T) )
# Could still add if the amount is negative and a round number but dont have
time for that now
View(tr[tr$geldautomat_or_auszahlung == F & tr$category_id == 10,])
```

Around 80 transactions from training data not caught by this rule, for this time and effort that is very ok.

Applying the rule for test data.

```
test$geldautomat_or_auszahlung <- ( grepl("geldautomat", test$transaction_type,
ignore.case=T) | grepl("auszahlung", test$transaction_type, ignore.case=T) )
test$category_id <- ifelse(test$geldautomat_or_auszahlung, 10, 27)
```

```
View(test[test$category_id == 10,])
nrow(test[test$category_id == 10,])
# [1] 14
```

Classified a whole of 14 operations which is disappointing.

We conclude that training and test data consist of differently distributed data so we have to change our approach a little.

Lots of classifications could be done this way but it will require quite a lot of manual work which we dont have time for in the scope of this assigment.

We take a look at the frequency of the words occurring in the usage field

```
usagestring <- paste(test$usage, collapse='')
words.list <- strsplit(usagestring, "\\W+", perl=TRUE)
words.vector <- unlist(words.list)
freq.list <- table(words.vector)
words <- head(sort(freq.list, decreasing=T), n=300)
words <- as.matrix(words)
words_test <- words[is.na(as.numeric(rownames(words))),]
```

Filtering out those that are only numeric since we dont have the time to look for patterns in those.

We get this list of frequencys of words occurring in the usage field:

SVWZ	NR	CRED	MREF	BIC	IBAN	VOM	Nr	KTO	EREF
2159	1679	1183	1067	703	701	491	488	460	428
GEHALT	EC	Ref	V	EINR	GA	EINZAHLUNG	KD	BERWEISUNG	vom
415	411	411	378	374	335	330	329	328	305
RE	END	SEPA	REF	F	BARGELDAUSZAHLUNG	PP	BELAST	DE	TAN
300	287	230	229	228	224	209	208	198	191
UHR	RG	r	R	BLZ	GMBH	S	Abrechnung	UND	ID
191	188	183	179	178	171	171	168	168	167
BERTRAG	KAUFUMSATZ	Beleg	FUER	M	f	Karten	AG	AN	A
164	164	158	158	156	155	152	149	148	147
TO	MAHNOLI	NR240015017	Lohn	BIS	Gehalt	LOHN	Verwendungszweck	AM	BANK
146	135	135	133	132	131	130	130	129	129
Mandatsref	NICHT	E	INTERNET	UM	ANGEGEBEN	N	B	BEITRAG	VON
122	121	118	116	114	112	112	111	110	110
2013EC	siehe	UEBERWEISUNG	COM	PAYPAL	PER	KONTO	Mandat	KU	Rechnung
109	109	108	104	104	104	101	101	100	98
Einreicher	ULTIMO	an	DA	ME1EC	ABSCHLAG	ABWA	LS	EUR	TA
97	97	97	95	94	93	93	93	92	89
BASISLASTSCHRIFT	NR989281028	ZUM	D	DE26ZZZ00000006194	Kontenclearing	Referenz	MIT	IHRE	MIETE
88	88	88	87	86	86	85	84	83	81
K	LU	ME0EC	AWV	BEACHTEN	BUNDESBANK	HOTLINE	MELDEPFLICHT	Ueberweisung	14EREF
80	80	80	77	77	77	77	77	77	76
RATE	RECHNUNG	Re	2014EREF	2014SVWZ	GmbH	Rg	BEI	Kunden	T
76	75	75	74	74	74	74	73	73	73
ABRECHNUNG	13EC	C	REISEBUERO	DANKE	NOTPROVIDED	CRS	ME0	TELEFON	EZUE
72	71	71	70	69	69	68	68	68	67
INV	WERTGARANTIE	VERLAG	ABO	LIEBE	WEG	CO	LENDER	U	ITUNES
67	67	65	63	63	63	61	61	61	60
KRT0000	SPAREN	ABBUCHUNG	von	Miete	G	IN	v	DE21ZZZ00000079131	ME2EC
60	60	58	58	57	56	56	56	55	55
VOLKSWAGEN	fuer	1EC	IHR	WWW	www	AUF	Kd	Rec	ZU
55	54	53	53	53	53	52	52	52	52
KONTONR	AMAZON	BERLIN	Vodafone	VIELEN	AIR	CARD	14SVWZ	DANK	SAGT
51	50	50	50	49	48	48	47	47	47
0387362825SPARRATE	EU	OTHR	PLC	STR	ng	I	KERPEN	KFZ	SCHIEFERSBURGER
46	46	45	45	45	45	44	44	44	44
1630035633Umbuchung	ABR	UST	IHNEN	Vertrag	hnungsnr	0387362817SPARRATE	CRN	ClickandBuy	LASTSCHRIFT
43	43	43	42	42	42	41	41	41	41
Rechnungsnr	2014EC	AB	AND	DIE	Dauerauftrag	EINZUG	VERTRAG	clickandbuy	RSB
41	40	40	40	40	40	40	40	40	39
ST	VERS	C93871589	P	SBT	STROM	2013SVWZ	AZ	Beitrag	DER
39	39	38	38	38	38	37	37	37	37
L	Rate								
37	37								

Full size link: http://i.imgur.com/EN9JDyy.png

Some of the words such reiseburo, Internet, telefone, Vodafone, Itunes, Miete etc. etc. can immediately help us categorize the transactions easily.

Some are harder such as Amazon where you can buy many different things such as electronics or books.

From a data-science perspective these simple cases are however not particularly interesting, although they are important.

Liebe can also be interpreted as many different things.
With more time would make the word frequency distribution also of-course case-insensitive

Now we have an idea of the word-frequency in the test data-set. The training and test data-set seem to be somewhat differently composed however.
Therefore it is a good idea to look at both data-sets to get an idea of what we can use to learn from.
At this point we therefore need to take a look at the training data-set which has the already classified items, to get an overview.

```
usagestring <- paste(tr$usage, collapse='')
words.list <- strsplit(usagestring, "\\W+", perl=TRUE)
words.vector <- unlist(words.list)
freq.list <- table(words.vector)
words <- head(sort(freq.list, decreasing=T), n=300)
words <- as.matrix(words)
words[is.na(as.numeric(rownames(words))),]
```

The word frequencies in the test-data are quite different from the training data:

```
> words[is.na(as.numeric(rownames(words))),]
     siehe          SVWZ              MREF             CRED                     NR                   VOM                IHRE                KONTONR                  EC
       893           839               733              726                    663                   395                 386                    347                 333
  RECHNUNG           Ref                PP           PAYPAL                 ITUNES                BELAST         RECHNUNGSNR          EURMASTERCARD                  DE
       313           297               295              256                    252                   245                 244                    207                 204
      EREF  VERTRAGSKONTO           VIELEN              IHR                  APPLE    DE93ZZZ00000078611              AMAZON                  STORE                RECH
       188           183               168              158                    155                   154                 152                    152                 150
        RE          VISA              DANK            KONTO          ZAHLUNGSBELEG             ABBUCHUNG            VODAFONE                    EUR                  ID
       149           138               134              134                    134                   131                 127                    122                 117
       COM           UHR             ME0EC              ABR AnlageEntgeltabrechnung      AnlageAbrechnung                  SB DE00020500061000000000                  EU
       116           111               109              106                    106                   103                 103                    102                 102
       BIC         KARTE                 V              HIT                   IBAN                     A            AnlageEC                    RGN DE97000000000142462
       100           100                97               96                     96                    93                  92                     83                  82
       GAA           WWW             SPKHB              END                   fuer                 DANKE  DE81ZZZ00000103174                    PER                SAGT
        82            81                80               74                     73                    71                  71                     71                  71
        Nr           CRN              15EC      CLICKANDBUY                 Mandat                   BIS                   R                  15UHR             BEITRAG
        70            68                67               67                     67                    66                  65                     64                  64
Einreicher          14EC            18UHR         INTERNET                     AG                  FUER                  ST       BASISLASTSCHRIFT          BERWEISUNG
        64            63                63               63                     62                    62                  62                     60                  60
     EUREC            IN             08UHR            57UHR                    ME5                     B                26EC                  30UHR               07UHR
        60            60                59               58                     57                    56                  55                     55                  54
         F          SEPA             09UHR            13UHR                  34UHR          42635401XXXX               52UHR                  58UHR               LEBEN
        54            54                53               53                     53                    53                  53                     53                  53
Mandatsref         SUMME  WOCHENABRECHNUNG            06UHR                  47UHR                   DRP                 KTO                  03UHR               37UHR
        53            53                53               52                     52                    52                  52                     51                  51
     43UHR           REF      T0010001B000              WIR                  11UHR                 14UHR               36UHR                  45UHR               49UHR
        51            51                51               51                     50                    50                  50                     50                  50
      ABWE            DO             19UHR            21UHR                  24UHR                 59UHR             KRANKEN                    SPK               25UHR
        50            50                49               49                     49                    49                  49                     49                  48
     31UHR         38UHR             48UHR            17UHR                  53UHR                     D               ME1EC          Vertragskonto               01UHR
        48            48                48               47                     47                    47                  47                     47                  46
     05UHR         23UHR             27UHR            29UHR                  41UHR                 54UHR                   E                    PAF             ZAHLUNG
        46            46                46               46                     46                    46                  46                     46                  46
     26UHR         33UHR             35UHR            46UHR             AnlageEREF                   STR               00UHR                  02UHR               04UHR
        45            45                45               45                     45                    45                  44                     44                  44
     44UHR         55UHR        ABRECHNUNG              BEI     DE95ZZZ00000029590                   ETT                   M                MARKTEC               SAGEN
        44            44                44               44                     44                    44                  42                     42                  42
 Zahlbeleg           vom             16UHR            NORDD                  28UHR                 51UHR            Anlage30                EINKAUF                GmbH
        42            42                41               41                     40                    40                  40                     40                  40
      Rech            TO             40UHR            42UHR                      C                 50UHR               56UHR                    DKB              ONLINE
        40            40                39               39                     39                    38                  38                     38                  38
       UND         20UHR             22UHR            ME2EC                      T                 10UHR               39UHR                      K                 ME0
        38            37                37               37                     37                    36                  36                     36                  36
    BERLIN            GL           MKTPLCE MONATSABRECHNUNG                    TEL                   TUR                TURK       2013ZAHLUNGSBELEG                  DB
        35            35                35               35                     35                    35                  35                     34                  34
Einzahlung  KUNDENNUMMER             ME7EC               NG                  12UHR                 32UHR                  AN                    BLZ                 CRS
        34            34                34               34                     33                    33                  33                     33                  33
       ME1          OGER            PAVILL           RECHNR                BERTRAG DE00020110020000000000              Gehalt
        33            33                33               33                     32                    32                  32
```

Full size link: http://i.imgur.com/rFhqF3X.png

We will focus on the most frequently occurring terms in the **test data** to focus on the features that potentially can classify most unclassified items.
That way we can stand a change of getting the most bang for the buck, that is classified items per used time/effort.

We start looking at the occurring words in the order of their frequency.

First question: Is the frequently occurring term SVWZ significant in determining the category?

```
tr$SVWZ <- grepl("svwz", tr$usage, ignore.case=T)
#View(tr[tr$SVWZ,])
table(tr[tr$SVWZ,]$category_id)
10  11  12  13  14  15  16  18  19  20  22  23  24  25  26
8   3   5 340   4  15  67  26  41   8  23  31 159  68 216

length(tr[tr$SVWZ,]$category_id)
#[1] 1014
```

So it yields a 30% chance of the category_id being 13 (Handy and internet)

Lets write a function for this:

```
findWord <- function(word, data) {
  data$foundWord <- grepl(word, data$usage, ignore.case=T)
  data
}
```

Now we need to look at the test data's most frequent terms and how frequent they were in the training data and how strong indication they give of the category. So we check how strong indication the most frequently occurring words give about the category and run them trough as a loop.

Taking 20 here as an arbitrary number, they seem to matter all the way to around 250.

```
for (i in 1:20) {
  term <- names(words_test[i])
  cat(" \n term: ", term, " \n ")
  data <- findWord(term, tr)
  distribution <- table(data[data$foundWord,]$category_id)
  cat("Distribution in training data: ")
  print(sort(distribution, decreasing=T))
  cat("Total occurrences: ", sum(distribution), " \n ")
}
```

Some of these terms have explanatory power based on the training data.

For example 2/3 of the transactions containing NR are classified in category 13 (Handy & Internet) in the training data.

```
term:  NR
Distribution in training data:
  13   26   18   24   10   22   23   25   11   20   12   16   17   14   19
1007  232  146   70   44   38   20   17    5    3    2    2    2    1    1
Total occurrences:  1590
```

Some other are quite interesting as well, the whole output from the 100 most frequent words are here:

https://docs.google.com/document/d/1f6bsHZp54K3rtSE4z-Vx49gIS8DW4Lq-euwflZ0OCFk/edit?usp=sharing

For example the last item Miete gives a strong indication of category 19 (Miete & Hypothek), 20 transactions of 26, and it possible the rest are wrongly categorized in the training data.

Barauszahlung sounds obvious but it is not present in training data so it would be a discretionary categorization.

Combining all the factors that have some explanatory power we can try forming a model to classify the transactions based on the training data.

We then start to engineer some features based on the term-frequency excercise we did previously.
The number of terms can be higher when the amount of data is increased.
Now we choose a low number since my time is a little scarce to choose the most significant ones and we dont want to overfit the model too much intentionally.

Loop to add columns to the data:

```
for (i in 1:40) {
  term <- names(words_test[i])
  data <- findWord(term, tr)
  tr <- cbind(tr, as.factor(data$foundWord))
  colnames(tr)[length(colnames(tr))] <- term
}
```

Some candidates to use for a RF classifier: transaction_type, amount
Some words that intuitively have explanatory power: gehalt miete lohn etc.
Unfortunately cannot be used for machine learning because training data does not have it: auszahlung.

Formatting the data and fitting the model:

```
dataformodel <- tr[,(16:length(colnames(tr)))]
dataformodel$amount <- tr$amount
dataformodel$contact_id <- tr$contact_id
dataformodel$category_id <- as.factor(tr$category_id)
```

So the data now has these fields:

```
cat(colnames(tr))
id booking_date effective_date gvcode usage additional transaction_type
user_bank_account_id category_id contact_bank_account_id transaction_group_id
id_hash balance amount contact_id SVWZ NR CRED MREF BIC IBAN VOM Nr KTO EREF
GEHALT EC Ref V EINR GA EINZAHLUNG KD BERWEISUNG vom
```

```
library(randomForest)
fit <- randomForest(category_id ~ . ,
                    data=dataformodel, importance=TRUE )
```

fit:

```
Call:
 randomForest(formula = category_id ~ ., data = dataformodel,      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 37.65%
Confusion matrix:
     10 11 12   13 14 15 16   17  18 19 20 21 22 23  24 25   26 class.error
10 2025  0  0   22  0  0  5  144   0  0  0  0  0  1   0  0  710      0.3034
11    1  5  0    1  0  0  0  202   0  0  0  0  0  0   0  0    6      0.9767
12    3  0 16    4  0  0  1  468   0  0  0  0  0  0   0  0   17      0.9686
13   74  0  0 1184  0  0  1   17   3  0  0  0  0  0  15  0  157      0.1840
14    0  0  0    1  0  0  0  219   0  0  0  0  0  0   2  0   13      1.0000
15    1  0  0    5  0  0  3    3   0  0  0  0  0  0   5  0    3      1.0000
16    9  0  1    1  0  0 55  119   0  0  0  0  0  0   0  0   30      0.7442
17   14  0  0    0  0  0  0 2233   2  0  0  0  0  0   0  0  306      0.1260
18    3  0  0   39  0  0  1   31 113  0  0  0  0  0   1  0   81      0.5799
19    1  0  0    5  0  0  1    3   0  3  0  0  0  0   9  0   28      0.9400
20   14  0  0    3  0  0  1    2   0  0  0  0  0  0   6  0    1      1.0000
21   10  0  0    0  0  0  0  125   0  0  0  0  0  0   0  0   18      1.0000
22    3  0  0   41  0  0  1   34   0  0  0  0  0  0   1  0   11      1.0000
23   17  0  0   20  0  0  1  423   4  0  0  0  0 12  12  0   14      0.9761
24   90  0  0   90  0  0  3  140   2  0  0  0  0  0 129  0  142      0.7836
25   61  0  0   20  0  0  0   51   1  0  0  0  0  0  24 11   23      0.9424
26  572  0  1   76  0  0  7  711   0  2  0  0  0  2  35  0 3499      0.2866
```

We get an OOB error rate of 37.65% on the training data-set.

The importance figures:

```
> importance(fit)
                     10     11       12     13       14     15       16     17      18         19      20        21      22      23     24     25     26 MeanDecreaseAccuracy MeanDecreaseGini
SVWZ              9.963  1.347 10.61018 21.867  6.00064  8.349 32.50900 11.247  4.5065   9.709425  2.8367  2.235999  5.7347 14.4406  7.784  9.738 19.594               27.214          77.2366
NR               10.404 11.066 10.66161 24.673  4.56966  3.556 11.20762 12.416 21.4838   4.299452  6.6573  2.240026 15.0148  9.7661 19.624 14.823 21.485               24.645         300.8206
CRED              5.658  2.005  5.29653 17.926 -0.08642  6.537  6.15224  5.269  2.3464   6.078654  0.8200  1.001002  8.5388  6.7272  4.033 11.424  9.033               17.232          29.2373
MREF              6.974  2.151  6.79259 19.912  2.35730  4.646  7.81735  6.739  5.5846   6.044570  2.1596  1.001002  8.3371  6.0242 13.311 12.535  8.864               19.414          43.5179
BIC               4.717  0.000  1.41582 12.483  0.00000 -1.001  5.39123  5.861 -2.1866   5.791857  0.0000  0.000000 -0.2325  1.0010 10.006  4.651  8.920               20.270          10.8671
IBAN              4.016  1.001  1.00819 10.176  1.00100  0.000  5.28583  6.197 -1.7020   4.646763  1.0010  1.001002 -0.8379 -1.7363 10.056  2.795  7.715               17.683           9.9471
VOM               6.200  8.990  4.01906 13.302  2.24387  6.469  6.39257 11.884  7.5080   1.001002  1.0010  1.001002  2.4395 11.3047  8.985  6.670 19.143               21.807          42.2713
Nr               11.076 10.084 10.57730 25.794  4.76967  2.869 11.07478 13.451 22.6002   5.601870  7.1957  1.724487 16.0075  8.1627 20.528 15.132 22.028               25.750         321.9968
KTO               4.210  0.000  1.00100  8.344  0.00000  1.001  1.41705  2.860 -1.1775   1.001002  0.0000  0.000000  0.0000  0.9815  6.941 24.970  6.826               22.871          12.9568
EREF              6.106  1.732  6.55786 26.263  3.01906  6.596  8.50392  6.030 11.1049 -11.814353  3.4155  1.707807  4.1107  9.9488 16.863 14.768  7.291               23.724          39.7510
GEHALT            2.103  0.000  7.69878 28.428  0.00000  0.000  1.99810  7.382  0.0000   1.414595  0.0000  0.000000  0.0000  0.0000  3.529  2.602 12.318               27.822          25.5438
EC               13.601 12.203 19.59524 25.066  5.54472  6.698 16.93956 36.200 21.0888   9.880888  7.4333  4.695439 12.9880 12.3332 35.178 23.363 38.811               43.237         616.0786
Ref               9.031  2.757 11.71856 15.397  4.55095  3.392  6.04926 12.821  8.8066   0.006076  1.6760  2.513265  7.9972  8.7943 10.767  8.793 21.145               20.054          66.5163
V                15.609 10.405 21.06079 23.632 10.02999  6.820 22.60039 28.717 18.3398  12.521699  8.6166 -0.005933  6.1803 19.4922 31.583 27.551 42.314               46.114         385.3144
EINR              1.339  0.000  0.01504  9.896  0.00000  0.000  0.00000  1.737 -3.4076   0.000000  1.0010  0.000000  2.7166  0.0000 12.916  0.000  0.199               14.759           5.7899
GA               10.534  1.414  3.78542 10.835  1.73354  0.000 -2.51596  7.802  7.3488   0.706340  7.0571  1.001002  2.6395  2.8907 10.125  8.795 12.982               17.281          18.2226
EINZAHLUNG        4.303  0.000  0.00000  3.484  0.00000  0.000  1.41635  5.378  1.7346   0.000000  0.0000  0.000000  0.0000  0.0000  1.416  1.001  7.199                8.947           4.8350
KD                7.078 13.587  2.34931 17.056  1.41603  1.415 -0.05819  6.708 35.0594  -1.001002  1.0010  0.000000  4.5066  8.1091 11.366  1.709 18.304               36.813          65.0730
BERWEISUNG        3.661  0.000  1.73332  7.635  0.00000  0.000  7.49460  2.068  7.3721   4.904709  0.0000  0.000000  1.4092  1.0010  2.579  5.481 12.666               16.677           9.6030
vom               6.495  8.628  3.88793 13.155  2.45807  8.512  5.35230 11.649  7.5029   2.424916  1.0010  1.001002  2.7473 10.5763  9.527  7.710 18.287               22.246          41.8956
RE                9.688 10.377 14.45274 45.321  5.98811  7.332 11.29309 13.601 26.2821   7.230835  6.8991  2.485818 13.1269 14.6051 17.594 18.129 16.789               35.451         277.9020
END               5.753  0.000  2.66258  7.386 11.24700  0.000  4.76376  2.940 10.1907   0.952559  0.0000  0.000000  1.0010  7.7305  4.276  2.246 -2.054               15.119          10.8443
SEPA              2.117  0.000  8.00157  8.685  0.00000  0.000  1.87558  3.418  6.7694   3.423659  1.0010  0.000000  7.6936  0.2718 13.723  5.002 15.519               23.543          16.4309
REF               9.219  1.057 13.03910 16.162  3.57043  3.753  9.33492 12.628  7.0341  -0.164330  0.5135  1.511794  8.2592  9.4459 10.945  9.048 20.280               20.596          69.0200
F                12.204  7.264 18.28790 15.961  7.65836  2.404  8.39955 14.550  9.5865   2.843666  3.4059  1.863030  9.6907  8.1027 24.687 13.081 19.312               32.657          86.2400
BARGELDAUSZAHLUNG 0.000  0.000  0.00000  0.000  0.00000  0.000  0.00000  0.000  0.0000   0.000000  0.0000  0.000000  0.0000  0.0000  0.000  0.000  6.234                6.227           0.7183
PP               11.777  1.939  4.40058 12.349  2.45429  1.723 -1.87045 12.050  5.3823   1.713510  1.7300  1.311754  3.0531  3.1844 26.774  7.510 12.202               23.740          53.7267
BELAST            6.578  3.775 21.14498 13.326  0.00000  1.001 -0.45753 31.232  9.7970   0.000000  1.0010  1.159654  2.9987 13.7225  8.127  4.426  9.747               36.554          60.2956
DE                8.190 -1.738  8.84415 27.917  4.13542  6.143  9.73909 11.246 28.5535   8.891806  8.1627  7.520626 10.4599 10.4622 19.511 16.130 14.231               36.959          81.0172
TAN               3.012  0.000  0.00000  7.138  0.00000  0.000  0.27011  6.311  1.0010  -0.374424  0.0000  0.000000  1.4166  0.0000  3.444 -1.417  3.300                9.169           3.4386
UHR              83.301  7.437 17.34251 24.589  6.38933  3.721 17.89286 19.278 22.5456   7.826570  7.0096  4.961948 13.6541 11.1788 36.664 27.070 33.230               68.415        1692.4658
RG                3.713  1.945  3.54403 18.594  2.46023  0.000  0.33086  6.543 23.7127   1.415084  6.6307  1.416423  3.6214  4.1908 10.512 16.137  9.992               27.560          42.8632
r                14.712  7.380  9.22536 12.672  5.65754  3.576  7.44899 16.048 12.2874   4.515461  6.6376 -2.021517  9.0113  3.9904 17.432 16.746  9.525               17.778         368.2736
R                16.171  6.426 10.06471 13.242  5.52270  4.324  7.89744 17.774 13.5814   5.499840  6.7069 -2.396844  9.8107  7.1173 18.308 18.331  9.177               19.651         401.7729
BLZ               6.224  0.000  1.00100  2.079  0.00000  0.000  0.00000  0.000  2.2438   0.000000  0.0000  0.000000  1.9768  1.0010  3.191  0.000 27.440               27.650          18.6337
GMBH              6.877  0.000  2.45831 12.955  1.41599  0.000  2.22814  7.753  0.5752   1.417051  8.6588  0.000000  1.4135  2.2411  5.290  2.650 19.519               22.894          18.4090
S                10.532  7.467 14.15277 22.605  9.45636  5.313 19.30034 26.910 16.8850   9.825258  9.1760 -4.701413  9.7145 10.4813 28.499 22.252 14.692               36.435         298.2182
Abrechnung       11.091  4.024  7.91040 17.491  3.81769  2.255  8.32112  9.209 13.0328   4.002899  3.1053  2.238574  5.7469 21.6843 14.829  8.543 12.555               17.528         171.9762
UND               9.099  1.417  2.00706 15.364  0.00000  1.001  6.99112  6.442 13.0797   1.001002  1.9763  1.001002  1.4155  5.3461  5.885 37.339 25.650               39.990          49.0740
ID                6.411  0.000  3.65044 19.097  1.00100  0.000  7.76107  7.942 11.9342   1.726822  1.0010  0.000000  7.3011  8.8740 12.972  4.444 13.527               28.914          25.0024
amount           16.384 11.241  8.51071 37.175  5.45028  6.062 27.84801 22.483 16.4540  14.041216 11.4052  5.429851 20.5717 22.3141 42.658 33.516 40.953               44.532         662.7938
contact_id       11.416  1.626 12.72864 27.827  9.68854  9.216 26.45805 20.274 18.1679  12.827648  9.8063  1.273042 11.1867 10.5763 38.856 25.903 38.209               49.294         295.1992
```

Applying the model on the test data-set

Adding the features to the test-data:

```
for (i in 1:20) {
  term <- names(words_test[i])
  data <- findWord(term, test)
  test <- cbind(test, as.factor(data$foundWord))
  colnames(test)[length(colnames(test))] <- term
}
```

Applying the model

```
test$predicted <- predict(fit, test)
```

```
head(test)
```

```
> head(test)
        id        booking_date      effective_date gvcode                                                                                  usage additional transaction_type user_bank_account_id category_id
2 30456710 2014-03-17 00:00:00 2014-03-17 00:00:00   NULL                                     QVZJ7Z/DE08600100700667567 01 RG 013-14 Casper Tour       NULL      ?BERWEISUNG              7350408          27
4 30456712 2014-03-13 00:00:00 2014-03-13 00:00:00   NULL                                        Referenz NOTPROVIDED Verwendungszweck RG 013-14       NULL       GUTSCHRIFT              7350408          27
5 30456713 2014-03-12 00:00:00 2014-03-12 00:00:00   NULL KBS 83132249 KRT0006/12.18 12.03 14.01 TA-NR. 146840 99084 Erfurt Anger 66-73 EC-CARD MIT PIN       NULL             NULL              7350408          27
6 30456714 2014-03-12 00:00:00 2014-03-12 00:00:00   NULL                                                  NYEF7Q/DE08600100700667567 01 sparen       NULL      ?BERWEISUNG              7350408          27
7 30456715 2014-03-12 00:00:00 2014-03-12 00:00:00   NULL                                        DDV3ZK/DE38760100850486451 8 54 RG 10-14 ICC ITB       NULL      ?BERWEISUNG              7350408          27
8 30456716 2014-03-11 00:00:00 2014-03-11 00:00:00   NULL                                          3WAR6K/DE08600100700667567 01 011-14 ICC ITB       NULL      ?BERWEISUNG              7350408          27
  contact_bank_account_id transaction_group_id                          id_hash balance  amount contact_id  SVWZ    NR  CRED  MREF   BIC  IBAN   VOM    Nr   KTO  EREF GEHALT    EC   Ref     V  EINR    GA EINZAHLUNG    KD BERWEISUNG
2                    NULL                 NULL 51bfa9a1488e2c504736f274581acbf9  8886.2 -1530.8    7666650 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE FALSE  TRUE FALSE FALSE      FALSE FALSE      FALSE
4                 7349571                 NULL 4504dbce51358ccba87f2882e775cdee 10426.0  9857.8    7666655 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE  TRUE  TRUE  TRUE FALSE FALSE      FALSE FALSE      FALSE
5                    NULL                 NULL a654495c22b4215d67c51b9097f8ee86   568.2 -4050.0    7667509 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  FALSE  TRUE FALSE FALSE FALSE FALSE      FALSE FALSE      FALSE
6                    NULL                 NULL 59c4122880692d9b820dc4931eba1baa  4618.2 -4000.0    7666650 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE FALSE FALSE FALSE FALSE      FALSE FALSE      FALSE
7                    NULL                 NULL 4703d1f06e701df80eb6ba2b887dc479  8618.2 -5916.8    7666652 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE FALSE  TRUE FALSE FALSE      FALSE FALSE      FALSE
8                    NULL                 NULL 8ac8d4d66618893e2afae5169ae44c46 14535.0  -687.4    7666650 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE FALSE FALSE FALSE FALSE      FALSE FALSE      FALSE
    vom predicted
2 FALSE        26
4 FALSE        26
5 FALSE        13
6 FALSE        10
7 FALSE        26
8 FALSE        10
```

Note: There are irregularities regarding the classifications in the training data and not enough time to do feature selection properly so this predictive step should be more looked at as an academic excercise.

Summary and conclusion

First we formed a simple rule for Bargeld which due to the distribution between test and training data did not yield many new classified items, but accuracy was high.

Then we formed a quick classification model with some features engineered and trained the model. Based on this quick excercise we got an OOB error rate of ca.38% in the training set which considering the time and effort spent has to be considered quite ok. The model should not in any case be considered production grade or ready for use because so many short-cuts were taken to save time and careful feature selection and testing has not been carried out and some irregularities in the training data affects the training of the model.

However, important to note:

- This was done quickly and many shortcuts were taken which when avoided would amount to better quality analysis in real life.
- Probably a healthy dose of overfitting is present, but it will be for later excercises to increase amount of data and do some careful feature selection and engineering.

- Some questions remain regarding classifications in the training data which would have to be clarified, eg. Miete http://i.imgur.com/jpD6SNY.png and ITunes http://i.imgur.com/XXJZB26.png these irregularities affect the training of the model in an unoptimal way.

To improve if there was more time:

- Many quick wins for the classification could be achieved by more manual work and simple rules, eg. barauszahlung and other more discretionary measures such as forming more rules like we did for **Bargeld**.
- Do proper feature selection and testing of significance (time-consuming process)
- Word frequency calculations should be case insensitive
- Use other characters also as separator, such as "+" for example
- New features to test in v.0.2:
    - Round number of transaction (eg. Bargeld is always round)
    - Sign of transaction +/-
    - Certain days more likely to trigger some transactions
    - Add the transaction_type as factor to the model
    - etc.etc.

Bank-account time series modeling and prediction

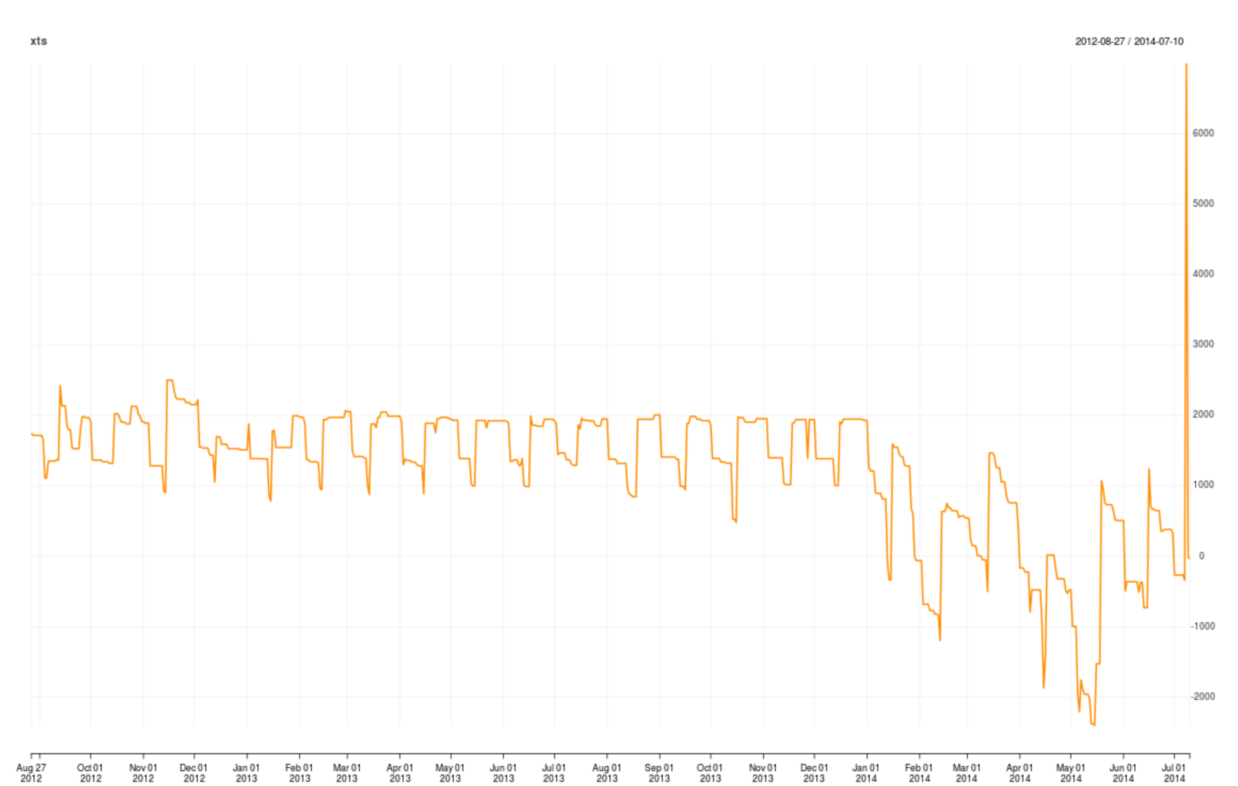
Now we take a look at a time-series of one persons bank-account. Our task is to model this data and predict it 90 days into the future.

We read in the data

```
ts <- read.csv(file="timeseries-forecasting.csv",head=F,sep=",")
xts <- as.xts(ts[,2], order.by=as.Date(ts[,1]))
colnames(xts) <- c("balance")
```

And chart it:

```
chart_Series(xts)
```



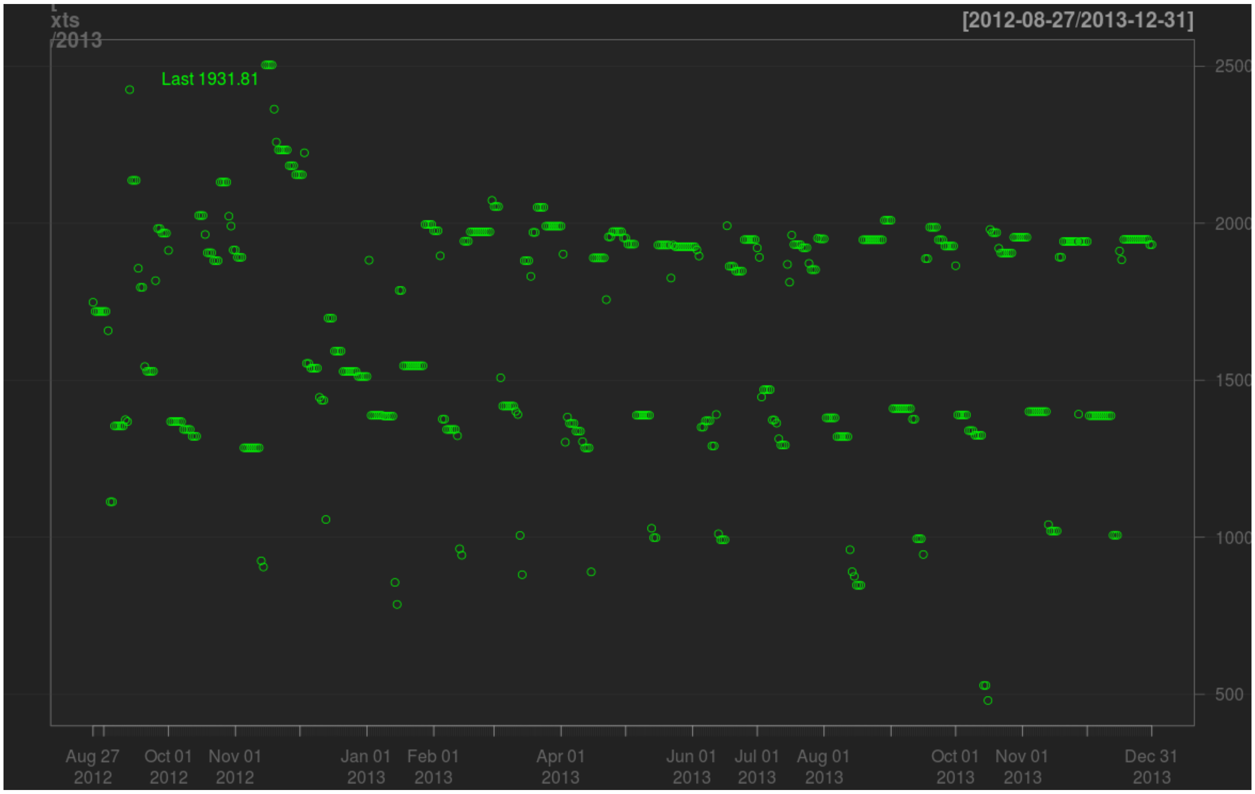
We can see that the data gets more volatile in 2014 so the task has been split up in 2 parts:

- Part 1: Pre 2014
- Part 2: Including 2014

Part 1: The less volatile period before 2014

We start by taking a closer look at the data and the separate data-points.

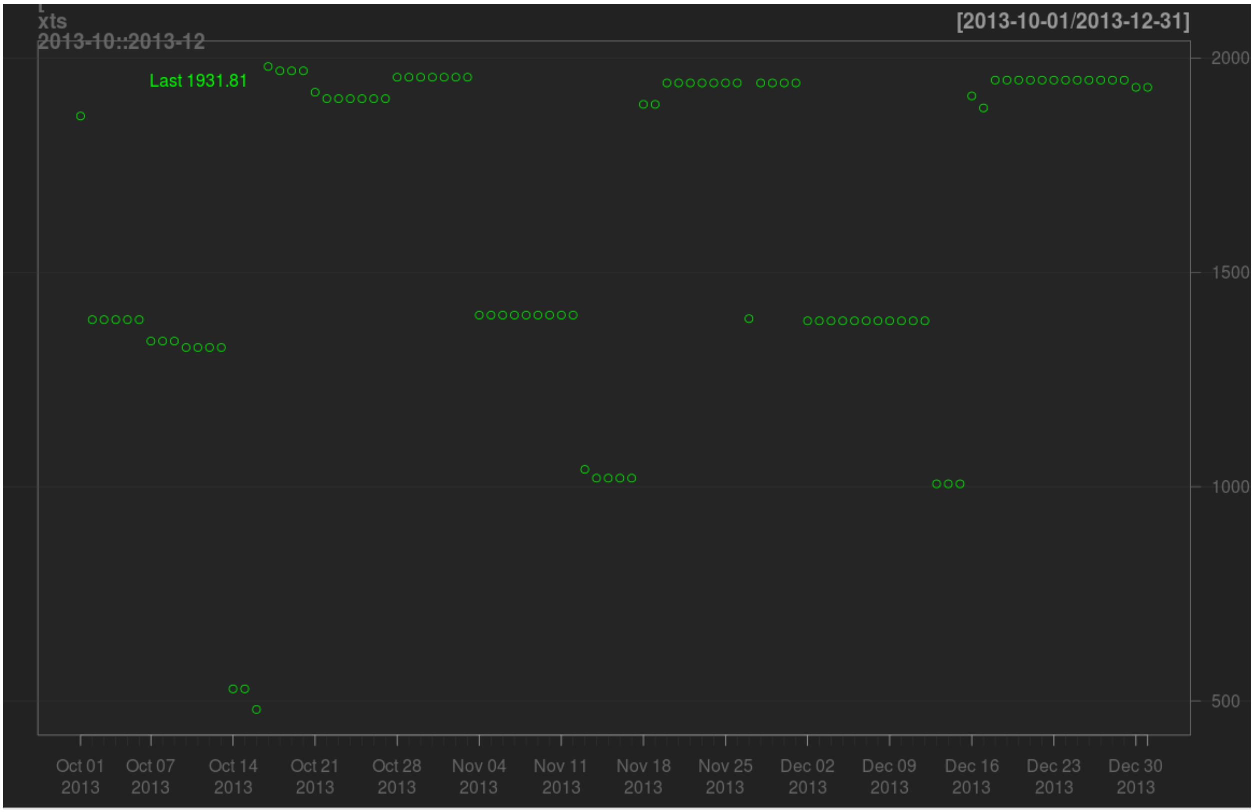
```
chartSeries(xts["/2013"], line.type="p")
```



Most of the data-points are around 1500 or 2000, the lower data-points around 1000 are fewer.

We take a closer look at the last 3 months of 2013.

```
chartSeries(xts["2013-10::2013-12"], line.type="p")
```



The closer look confirms the earlier suspicion about distribution.

Lets see how many datapoints we have per period:

```
nrow(xts["2013"])
[1] 365
```

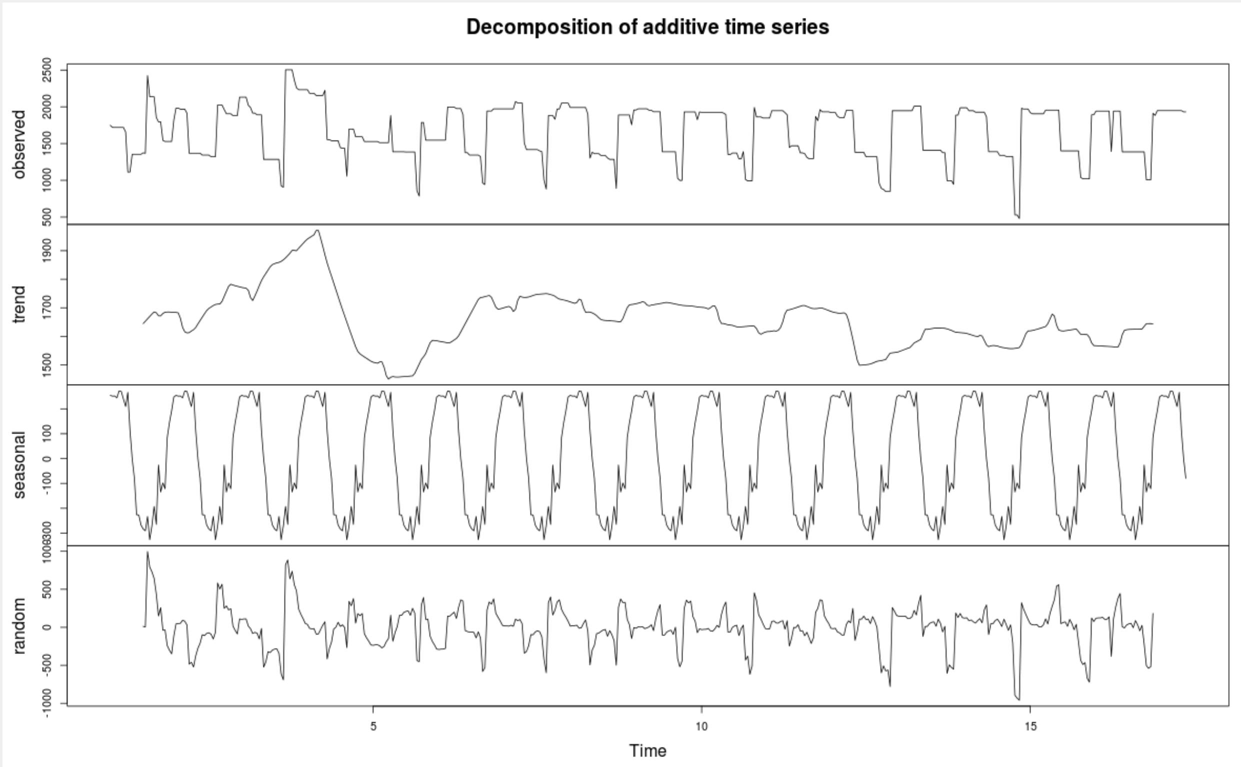
We have measurements for 365 days per year.

Splitting data into first part until Jan 2014

```
xts1 <- xts["/2013"]
```

We look at this from a monthly perspective and decompose the timeseries.

```
y <- ts(as.vector(xts1), frequency=30)
plot(decompose(y))
```



This quite nicely captures the seasonal element of the timeseries so we continue looking at it from a monthly perspective.

ARIMA

We fit an ARIMA model to the series:

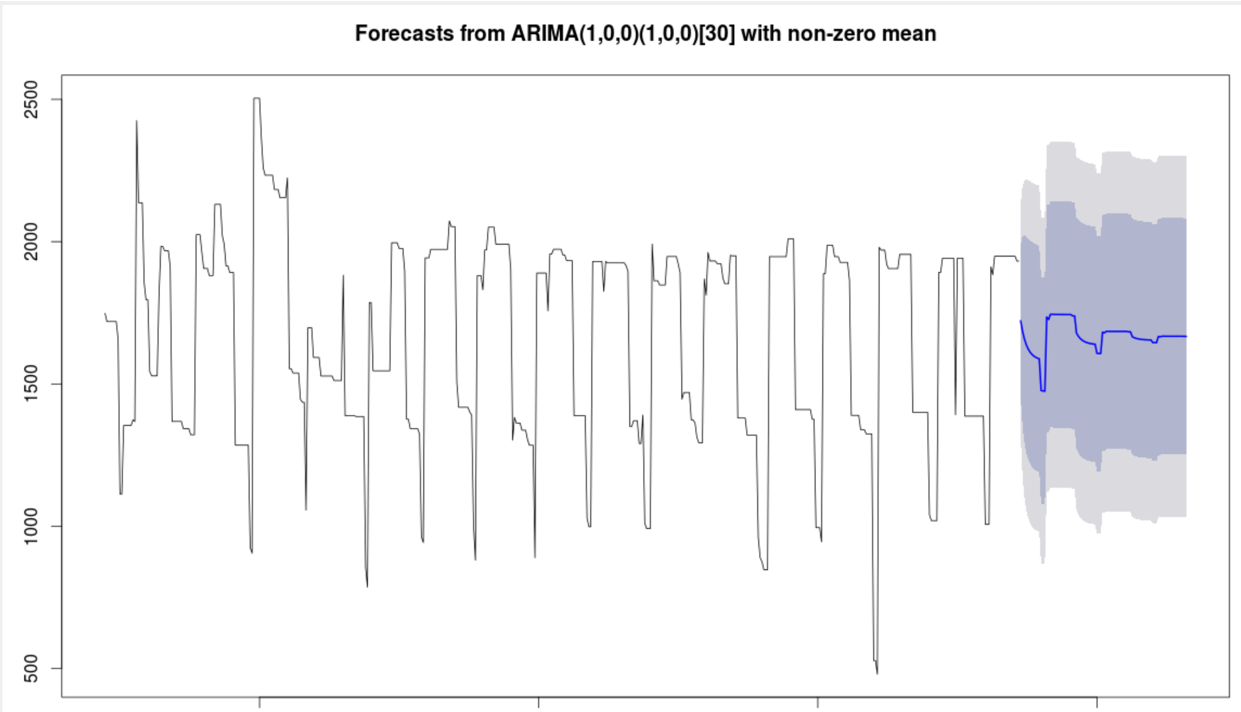
```
library(forecast)
ts2 <-ts(as.vector(xts1),start=c(2012,08),frequency=30)
#modArima <- auto.arima(ts2, D=NA, max.P = 5, max.Q = 5, trace=T)
modArima <- auto.arima(ts2, trace=T)
```

```
ARIMA(2,0,2)(1,0,1)[30] with non-zero mean : 6666
ARIMA(0,0,0) with non-zero mean : 7189
ARIMA(1,0,0)(1,0,0)[30] with non-zero mean : 6661
ARIMA(0,0,1)(0,0,1)[30] with non-zero mean : 6805
ARIMA(1,0,0) with non-zero mean : 6701
ARIMA(1,0,0)(2,0,0)[30] with non-zero mean : 6670
ARIMA(1,0,0)(1,0,1)[30] with non-zero mean : 6662
ARIMA(1,0,0)(2,0,1)[30] with non-zero mean : 6664
ARIMA(0,0,0)(1,0,0)[30] with non-zero mean : 6991
ARIMA(2,0,0)(1,0,0)[30] with non-zero mean : 6664
ARIMA(1,0,1)(1,0,0)[30] with non-zero mean : 6663
```

```
 ARIMA(2,0,1)(1,0,0)[30] with non-zero mean : 6664
 ARIMA(1,0,0)(1,0,0)[30] with zero mean     : 6721

 Best model: ARIMA(1,0,0)(1,0,0)[30] with non-zero mean
```

We plot the forecast plot(forecast(modArima, h=90))



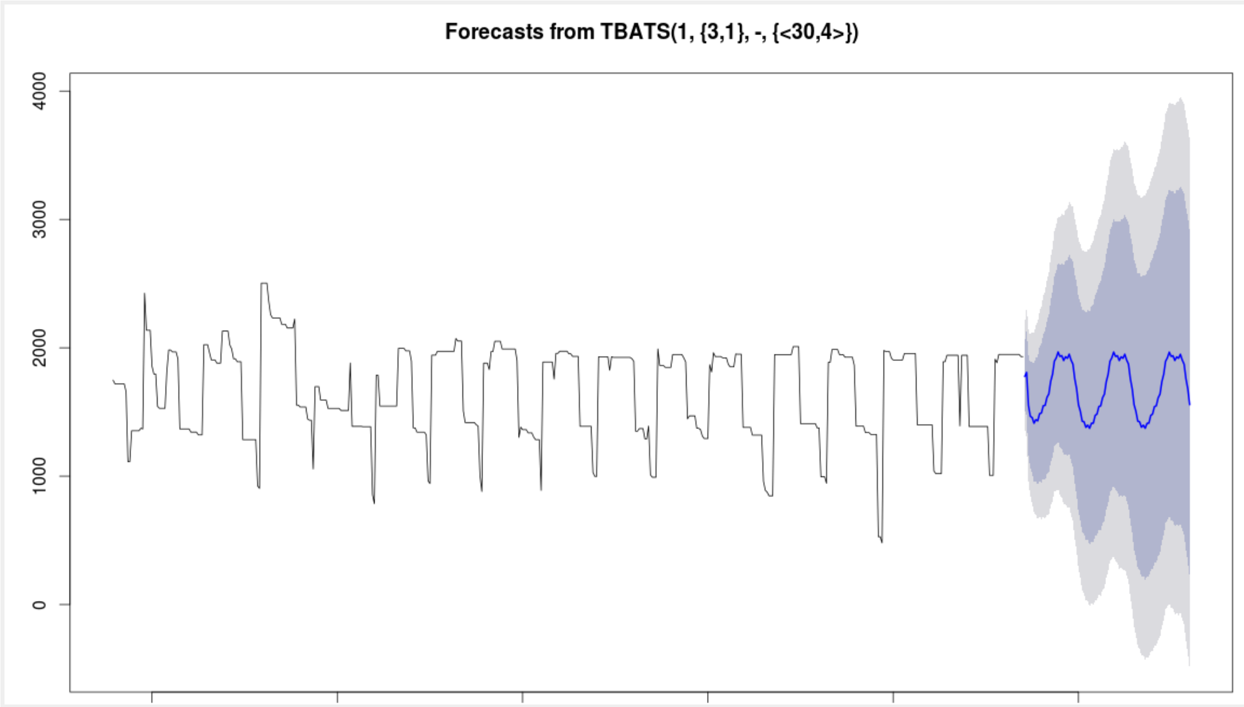
It looks good initially but then reverts to the mean as ARIMA models do.

TBATS

Out of curiosity we take a look at fitting a tbats model to the data. Tbats stands for (Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components).

```
ts2 <- as.zoo(xts1)
fit <- tbats(ts2, seasonal.periods=30)
```

And we again plot the forecast plot(forecast(fit, h=90))

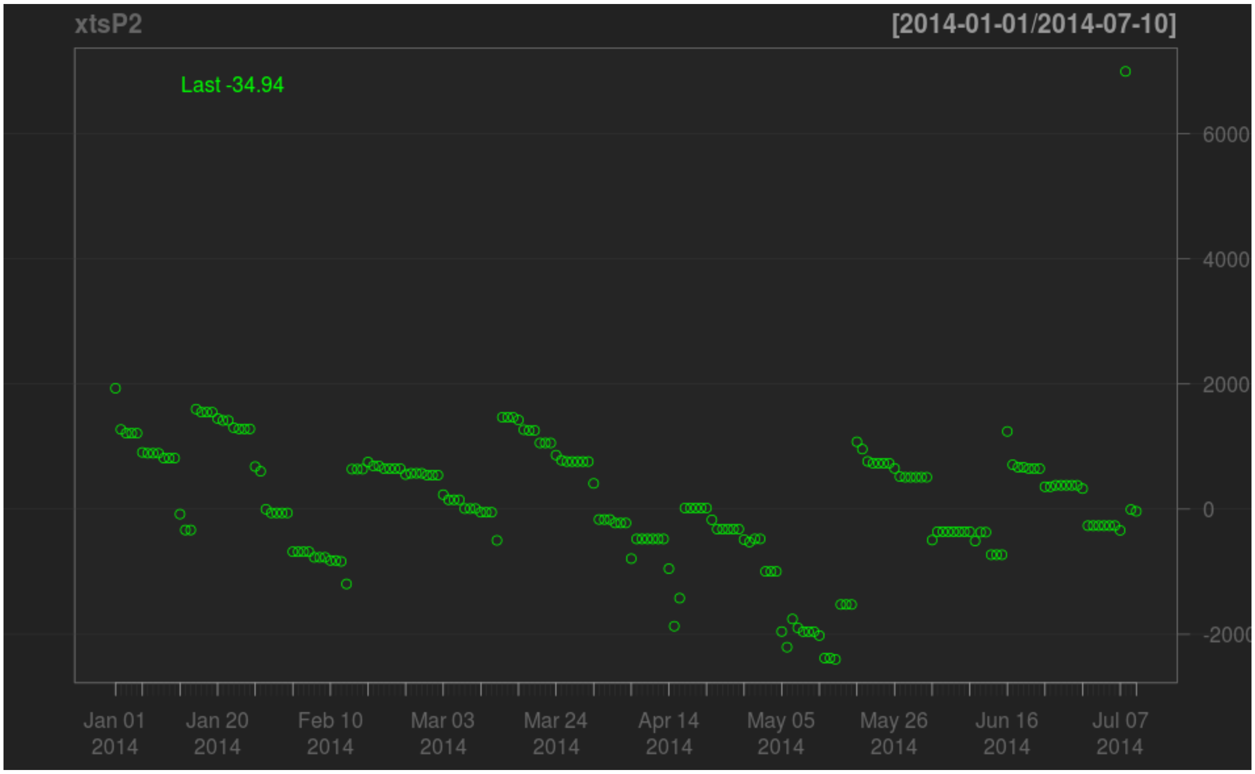


Yes, this model captures the seasonality better.

Part 2: The more volatile period.

We get the data for this period and chart it with the data-points visible.

```
xtsP2 <- xts["2014/"]
chartSeries(xtsP2, line.type="p")
```

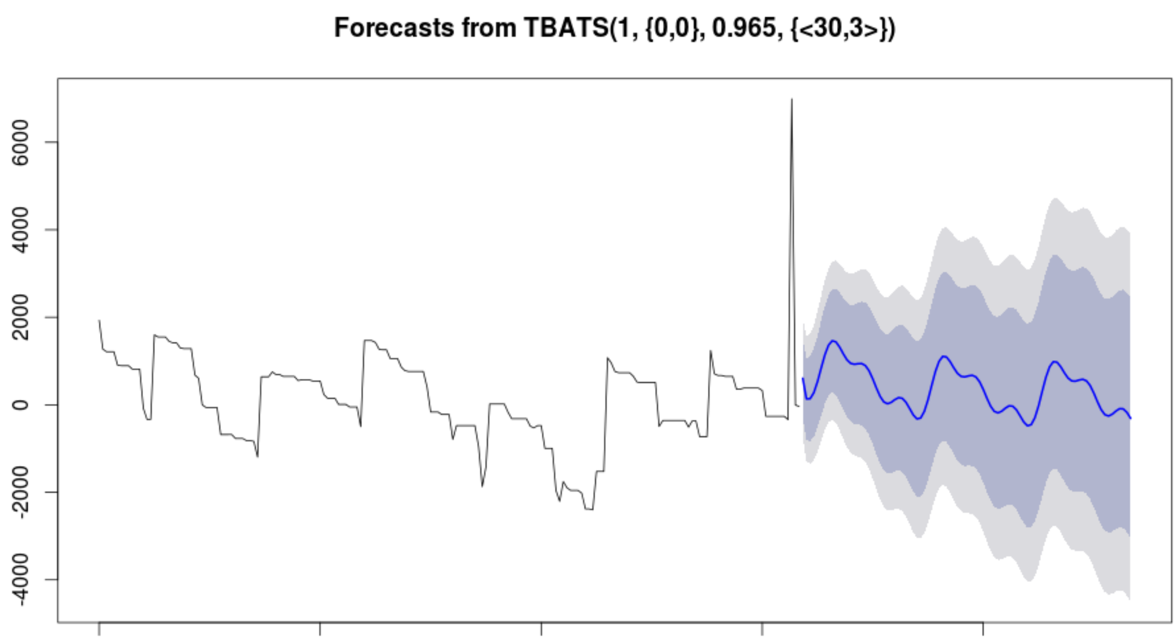


Interesting that it is only one datapoint above 6000.

We proceed by fitting a tbats model.

```
tsP2 <-ts(as.vector(xtsP2),start=c(2014,01),frequency=30)
fit <- tbats(tsP2) #, seasonal.periods=30)
```

And plotting the forecast `plot(forecast(fit, h=90))`



Assuming this was real data it is interesting to see that the spending increased before the large account inflows. It would be interesting to study if that holds true on a wider population.