Solution for MEM Assignment r2

Wenbo CHEN

Question #1: BigBangTheory. (Attached Data: BigBangTheory)

a. The minimum is 13.3, the maximum number of viewers is 16.5

b. Mean=15.04; median = 15; mode = 13.6

c. Q1 = 14.1; Q3 = 16

d. From the plot below, we cannot find any trend between 2011-2012 season.

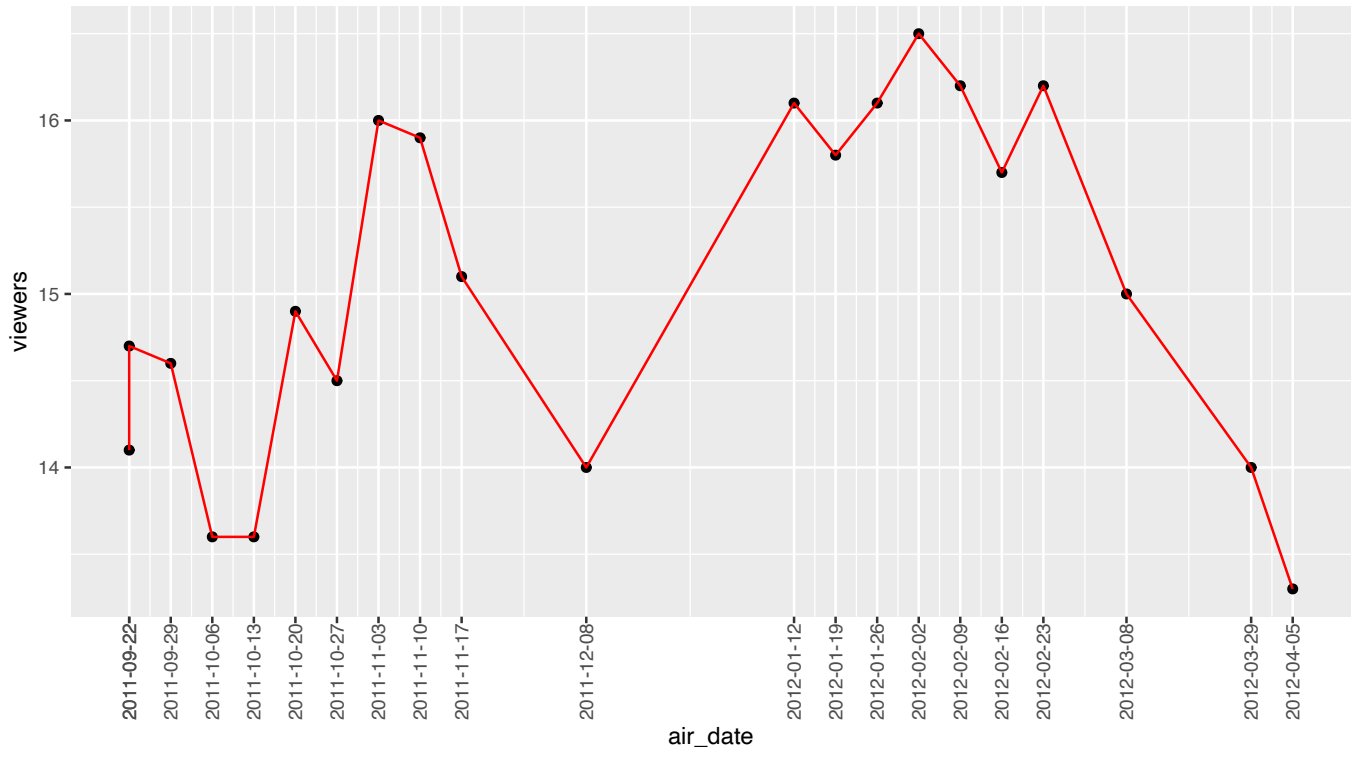


图 1: Plot between date and viewers

Question #2: NBAPlayerPts. (Attached Data: NBAPlayerPts)

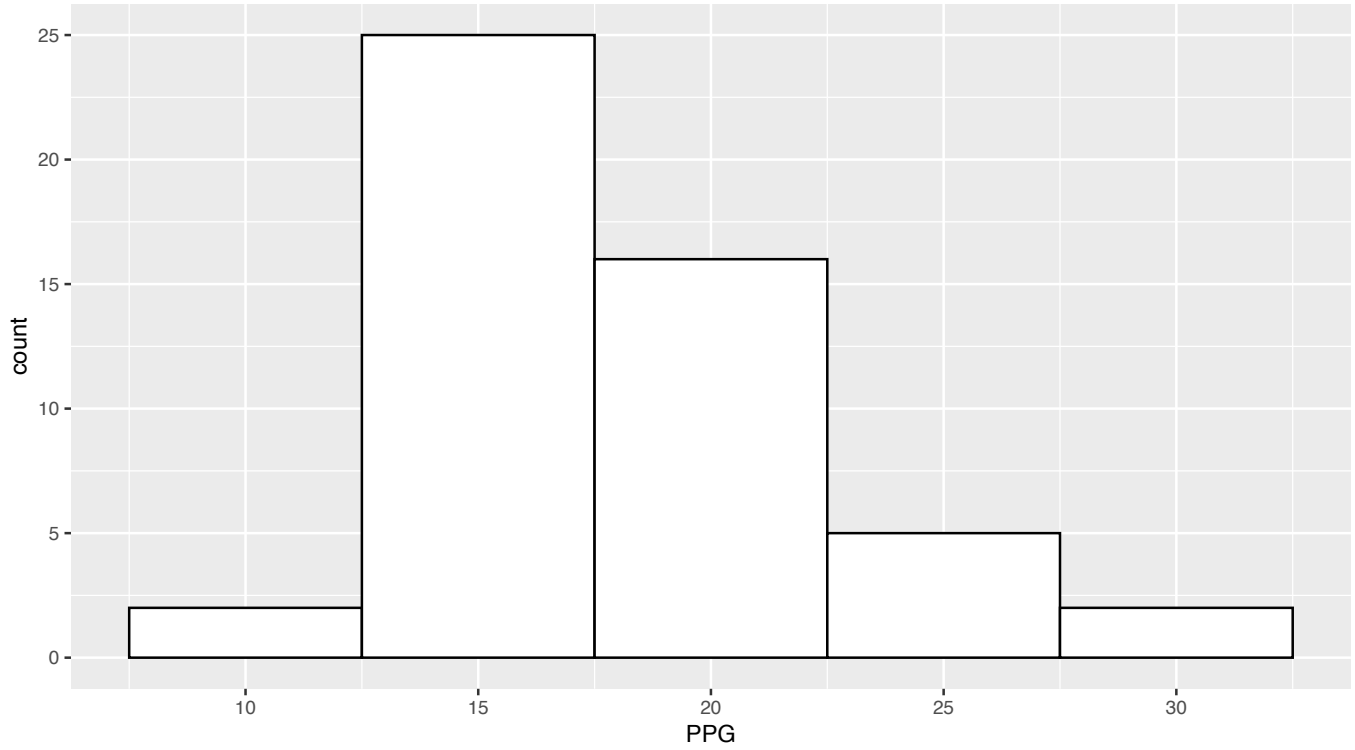
```
## [10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
##    0.02    0.10    0.22    0.62    0.78    0.86    0.90    0.90    0.96    1.00
```

a. The frequency distribution: 1, 4, 6, 20, 8, 4, 2, 0, 3, 2

b. Relative frequency: 0.02, 0.08, 0.12, 0.4, 0.16, 0.08, 0.04, 0, 0.06, 0.04

c. Cumulative percent frequency distribution: \n 0.02, 0.1, 0.22, 0.62, 0.78, 0.86, 0.9, 0.9, 0.96, 1

d. The histogram is as follow:



e. It seems skewed rightly, for it has a long tail to the right.

f. 1-78% = 22%.

Question #3

a. The sample size is 625.

b. The sample size is large, so the distribution of x-bar is normal distribution. The probability that the point estimate was within ±25 of the population mean is: 0.79.

skim_type	skim_variable	n_missing	complete_rate	factor.ordered	factor.n_unique	factor.top_counts	nu
factor	gender	0	1	FALSE	2	Mal: 229, Fem: 181	
factor	real_estate	0	1	FALSE	2	No: 229, Yes: 181	
factor	has_broadband	0	1	FALSE	2	Yes: 256, No: 154	
factor	have_children	0	1	FALSE	2	Yes: 219, No: 191	
numeric	age	0	1	NA	NA	NA	
numeric	investments	0	1	NA	NA	NA	
numeric	num_trans	0	1	NA	NA	NA	
numeric	income	0	1	NA	NA	NA	

Question #4

a. the descriptive statistics.

b. 95% confidence intervals of the mean age and household income.

```
## [1] 30 31
## attr(,"conf.level")
## [1] 0.95
```

```
## [1] 71079 77840
## attr(,"conf.level")
## [1] 0.95
```

c. We can conclude with 95% confidence that the mean age of subscribers to Young Professional is between 29.72 and 30.50 years of age. And, we can conclude with 95% confidence that the mean household income of subscribers is between $71,079 and $77,840.

```
## # A tibble: 1 x 6
##   statistic chisq_df     p_value alternative lower_ci upper_ci
##       <dbl>    <int>       <dbl> <chr>          <dbl>    <dbl>
## 1      24.9        1 0.000000610 two.sided      0.575    0.671
```

```
## # A tibble: 1 x 6
##   statistic chisq_df p_value alternative lower_ci upper_ci
##       <dbl>    <int>   <dbl> <chr>          <dbl>    <dbl>
## 1      1.78        1   0.182 two.sided      0.485    0.583
```

d. Yes. Young Professional should be a good advertising outlet for online brokers. We see that most of the subscribers have financial investments exclusive of their home (the mean amount is $28,538) and some of

them have a substantial amount of investments. (Several have over $100,000 of investments). Another factor to consider is the number of stock, bond, and mutual fund transactions. The mean number is approximately 6 per year and several subscribers make significantly more transactions than that. Finally a large proportion of subscribers have broadband access (the sample proportion is .6244) and this makes them more likely to do business with an online broker.

e. Yes,The survey results allow us to estimate that the mean age of subscribers is 30.12 years and that 53.41% of subscribers have children. Given the age of subscribers, it is reasonable to assume that their children are young. Thus, we conclude that subscribers to Young Professional would be a good target market for companies selling educational software and computer games for young children.

f. A variety of answers are possible here. But, from the survey results, it seems clear that articles about investing would have appeal to many readers. Articles about real estate and architecture would probably appeal to those subscribers planning to make a real estate purchase. Technology related articles would probably appeal to readers as well as an occasional article on parenting and child care.

Question #5: Quality Associate, Inc. (Attached Data: Quality)

a. the p_value is as follows:

```
##     s1     s2     s3     s4
## 0.2810 0.4547 0.0038 0.0339
```

a. Also, you can use interval to test the hypothesis

```
## [1] 12 12
```

```
## $s1
## [1] 12
##
## $s2
## [1] 12
##
## $s3
## [1] 12
##
## $s4
## [1] 12
```

b. s.d.

```
## $s1
## [1] 0.22
##
## $s2
## [1] 0.22
##
## $s3
## [1] 0.21
##
## $s4
## [1] 0.21
```

It's reasonable to assume the sd is 0.21.

c.

```
## [1] 12 12
```

d. e.g., $s.g. = 0.05$

```
## [1] 12 12
```

with the increase of significant level, type I error will increase.

Question #6

a. point estimate

```
## [1] 0.35
```

```
## [1] 0.47
```

b. 95% confidence interval

- The interval is -0.22, -0.01.

c. Yes. The interval doesn't include Zero. Which means we should reject the equality hypothesis.

skim_type	skim_variable	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	num
numeric	Current	0	1	75	3.9	65	72	
numeric	Proposed	0	1	75	2.5	69	74	

Question #7

a.

b.

```
##
##  Welch Two Sample t-test
##
## data:  data_q7$Current and data_q7$Proposed
## t = -0.6, df = 102, p-value = 0.5
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.55  0.83
## sample estimates:
## mean of x mean of y
##        75        75
```

- In 0.05 significant level, there is no difference between the two groups.

c.

```
## $Current
## [1] 3.9
##
## $Proposed
## [1] 2.5

## $Current
## [1] 16
##
## $Proposed
## [1] 6.3
```

```
##
##  F test to compare two variances
##
## data:  data_q7$Current and data_q7$Proposed
## F = 2, num df = 60, denom df = 60, p-value = 0.0006
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.5 4.1
## sample estimates:
## ratio of variances
##                2.5
```

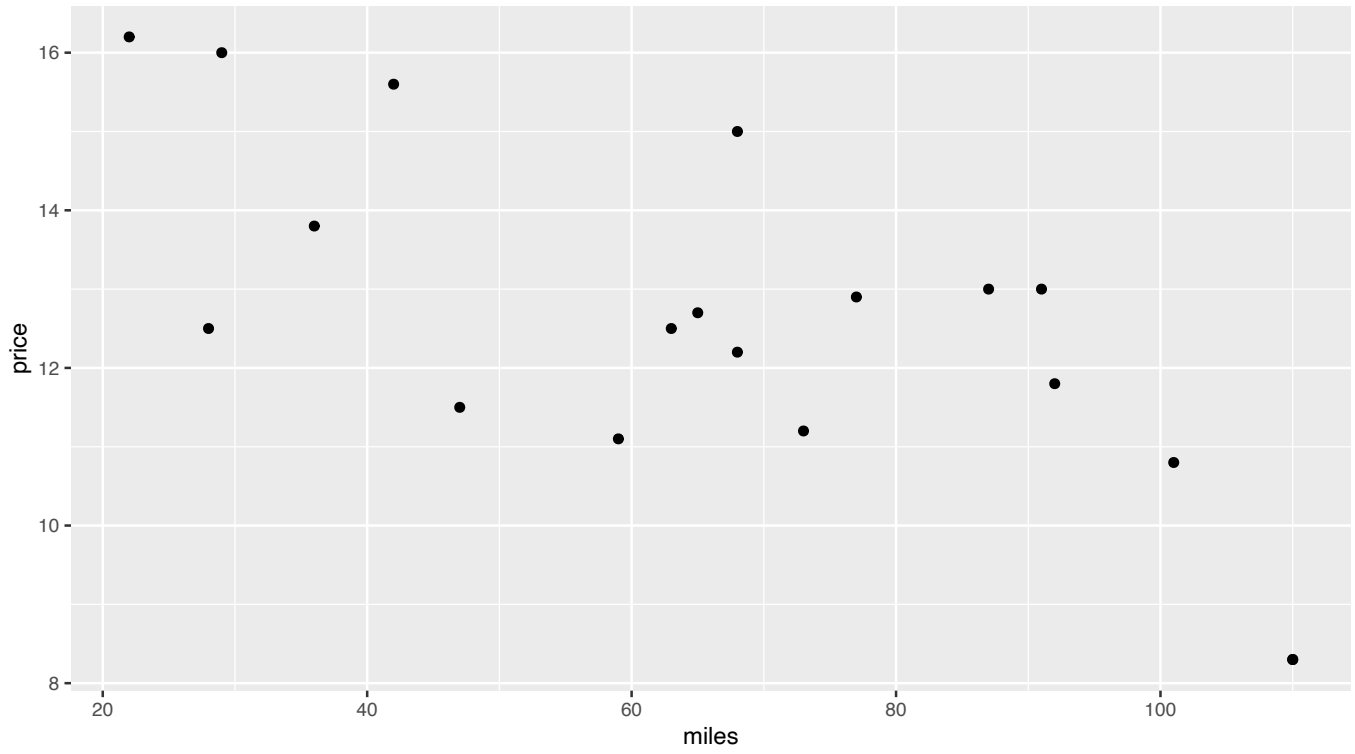
- conclusion: the sd, or variance is different. The Current method has larger variance.

d.Based on the data available, the proposed method is preferred. The two methods are very close in terms of mean completion times with the 95% confidence interval of the difference being -1.55 to 0.83 hours. However, the proposed method has a significantly lower variance. Under the proposed method, students are more likely to complete the training in approximately the same amount of time. There should be less chance of faster students waiting for slower students to complete the training.

e. Before making a final decision, we recommend that data be collected on the amount of learning under the two methods. The time data favors switching to the proposed method. However, is the quality of the training with the proposed method the same or better than the quality of the training with the current method? Both groups could be given an examination at the end of the training program. Analysis of the examination scores would determine if the programs were similar or different in terms of the amount of learning provided by the programs. This analysis should be made prior to the final decision to switch to the proposed method.

Question #8

a. The plot is as follows:



b. There appears to be a negative relationship between the two variables that can be approximated by a straight line. An argument could also be made that the relationship is perhaps curvilinear because at some point a car has so many miles that its value becomes very small.

c.

```
##
## Call:
## lm(formula = price ~ miles, data = data_q8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3241 -1.3419  0.0506  1.1290  2.5269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.4698     0.9488   17.36    3e-12 ***
## miles        -0.0588     0.0132   -4.46  0.00035 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.5 on 17 degrees of freedom
## Multiple R-squared:  0.539,  Adjusted R-squared:  0.512
## F-statistic: 19.8 on 1 and 17 DF,  p-value: 0.000348
```

- Regression Equation;

$$Price = 16.470 - 0.059 * miles$$

d. Significant relationship: $p - value = 0.000348 < \alpha = .05$

e. $R^2 = .5387$; A reasonably good fit considering that the condition of the car is also an important factor in what the price is.

f. The slope of the estimated regression equation is -.059. Thus, a one-unit increase in the value of x coincides with a decrease in the value of y equal to .059. Because the data were recorded in thousands, every additional 1000 miles on the car's odometer will result in a $59.0 decrease in the predicted price.

g. The predicted price for a 2007 Camry with 60,000 miles is $= 16.47 - .0588(60) = 12.942$ or $\$12,942$. Because of other factors, such as condition and whether the seller is a private party or a dealer, this is probably not the price you would offer for the car. But, it should be a good starting point in figuring out what to offer the seller.

Question #9

a. Visual exploration on the comparison between churn=0 and churn = 1.

```
## Rows: 6,347
## Columns: 13
## $ id           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ churn        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ happy_index  <dbl> 0, 62, 0, 231, 43, 138, 180, 116, 78, 78, 91, 40, 215, 0,~
## $ chg_hi       <dbl> 0, 4, 0, 1, -1, -10, -5, -11, -7, -37, -1, 14, 15, 0, 63,~
## $ support      <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ chg_supprt   <dbl> 0, 0, 0, -1, 0, 0, 1, 0, -2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ priority     <dbl> 0, 0, 0, 3, 0, 0, 3, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, ~
## $ chg_priority <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ log_in_fre   <dbl> 0, 0, 0, 167, 0, 43, 13, 0, -9, -7, 14, 0, 71, 0, 5, 0, 4~
## $ chg_blog_fre <dbl> 0, 0, 0, -8, 0, 0, -1, 0, 1, 0, 3, 0, 9, 0, 1, 0, 0, 0, 6~
## $ chg_vis      <dbl> 0, -16, 0, 21996, 9, -33, 907, 38, 0, 30, 0, 15, 8658, 0,~
## $ y_age        <dbl> 72, 72, 60, 68, 62, 63, 62, 51, 61, 61, 58, 61, 62, 62, 6~
## $ chg_interval <dbl> 33, 33, 33, 2, 33, 2, 2, 8, 9, 16, 2, 33, 2, 33, 2, 33, 3~
```

churn	happy_index	chg_hi	support	chg_supprt	priority	chg_priority	log_in_fre	chg_blog_fre	chg_vis
0	89	5.5	0.72	-0.01	0.83	0.03	16.1	0.17	107
1	63	-3.7	0.37	0.04	0.50	-0.02	8.1	-0.10	-96

variable	estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method
chg_blog_fre	0.27	0.17	-0.10	2.53	0.01	696	0.06	0.49	Welch Two San
chg_hi	9.27	5.53	-3.74	5.78	0.00	366	6.12	12.42	Welch Two San
chg_interval	-4.97	3.51	8.49	-4.10	0.00	346	-7.36	-2.59	Welch Two San
chg_priority	0.05	0.03	-0.02	0.64	0.52	364	-0.10	0.20	Welch Two San
chg_supprt	-0.05	-0.01	0.04	-0.63	0.53	407	-0.19	0.10	Welch Two San
chg_vis	202.38	106.61	-95.77	1.91	0.06	448	-5.46	410.22	Welch Two San
happy_index	25.33	88.61	63.27	7.62	0.00	369	18.80	31.87	Welch Two San
log_in_fre	8.08	16.14	8.06	3.57	0.00	363	3.63	12.53	Welch Two San
priority	0.33	0.83	0.50	5.14	0.00	373	0.20	0.46	Welch Two San
support	0.35	0.72	0.37	5.51	0.00	419	0.23	0.48	Welch Two San
y_age	-1.53	18.82	20.35	-2.98	0.00	380	-2.55	-0.52	Welch Two San

We find:

- There are differences among all the 11 indicators between those churn and not churn.
- But whether these differences are significant, we need to test.

b. using `t.test` to check whether the differences are significant.

From the table, we can get conclusion that:

- Except `chg-priority` and `chg_supprt`, all the other differences are significant.

c. d. Using logit regression to estimate the regression function, and then use the regression function to predict.

```
##
## Call:
## glm(formula = churn ~ chg_blog_fre + chg_hi + chg_interval +
##     chg_vis + happy_index + log_in_fre + priority + support +
##     y_age, family = binomial(link = "logit"), data = we_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
```

```
## -0.951  -0.354  -0.294  -0.235   2.928
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8744187  0.1214855  -23.66  < 2e-16 ***
## chg_blog_fre -0.0000236  0.0207961    0.00    0.999
## chg_hi       -0.0095009  0.0024239   -3.92  8.9e-05 ***
## chg_interval  0.0170011  0.0042767    3.98  7.0e-05 ***
## chg_vis      -0.0001171  0.0000407   -2.88    0.004 **
## happy_index  -0.0052250  0.0011610   -4.50  6.8e-06 ***
## log_in_fre    0.0009104  0.0019523    0.47    0.641
## priority     -0.0372737  0.0751390   -0.50    0.620
## support      -0.0352242  0.0743801   -0.47    0.636
## y_age         0.0141828  0.0052602    2.70    0.007 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2445.9  on 6337  degrees of freedom
## AIC: 2466
##
## Number of Fisher Scoring iterations: 6

## chg_blog_fre       chg_hi chg_interval      chg_vis  happy_index   log_in_fre
##          1.1          1.2          1.2          1.0          1.5          1.3
##     priority      support        y_age
##          2.1          2.2          1.2
```

Predict the churn probability.

Here, we provide the table on the 30 most likely churn customers. Customer retention should be taken to these customers.

id	churn	happy_index	chg_hi	support	chg_supprt	priority	chg_priority	log_in_fre	chg_blog_fre	ch
2287	0	227	7	5	5	2.8	2.8	11	-4	
109	0	0	-125	0	0	0.0	0.0	-8	0	
1971	0	0	-113	0	0	0.0	0.0	-23	0	
2025	0	18	-15	0	0	0.0	0.0	0	0	
1	0	0	0	0	0	0.0	0.0	0	0	
929	0	123	35	0	0	0.0	0.0	7	1	
2076	0	29	-69	0	0	0.0	0.0	0	0	
76	0	1	-70	0	0	0.0	0.0	-7	-1	
14	0	0	0	0	0	0.0	0.0	0	0	
18	0	0	0	0	0	0.0	0.0	0	0	
3	0	0	0	0	0	0.0	0.0	0	0	
2244	0	16	-38	0	0	0.0	0.0	0	0	
21	0	0	0	0	0	0.0	0.0	0	0	
1287	0	24	-72	0	-1	0.0	-3.0	-6	-1	
1929	0	7	-40	0	0	0.0	0.0	0	0	
1459	0	0	-22	0	0	0.0	0.0	0	0	
51	0	1	0	0	0	0.0	0.0	0	0	
128	0	31	-26	0	0	0.0	0.0	0	0	
183	0	0	-17	0	0	0.0	0.0	-1	0	
59	0	0	0	0	0	0.0	0.0	0	0	
55	0	3	0	0	0	0.0	0.0	0	0	
121	0	0	0	0	0	0.0	0.0	0	0	
2240	0	0	-15	0	0	0.0	0.0	0	0	
1520	0	0	-67	0	0	0.0	0.0	-4	0	
2599	0	7	-30	0	0	0.0	0.0	0	0	
1236	0	0	-35	0	0	0.0	0.0	0	0	
137	0	0	0	0	0	0.0	0.0	0	0	
1862	0	0	-27	0	0	0.0	0.0	0	0	
2080	0	4	-25	0	0	0.0	0.0	0	0	
1143	0	0	-17	0	0	0.0	0.0	0	0	