Spreadsheet Project Report

Kim Tsang

The completed project can be found here:

Data

There were two data sets. Data was given to students by the university of Koblenz-Landau, Germany. One was "translation data" and the other one was "payment data". The raw translation data had a total of 1036 records. Upon an initial inspection for data validity, a portion of the data was duplicated, misspelled, and written in a different format than others. After data cleaning, the data set returned 1008 unique records. Payment data on the other hand had a total 31 records. The data set did not have duplicated nor misspelled data, however data with a wrong format was found. After cleaning, the data set still contained 31 records and all entries were changed to the format of currency.

Changing the payment data from USD to Euro:

The conversion rates were found on google based on the date which the payment was made. As an example, the conversion rate on December 1st, 2022 for exchanging USD to Euro was 1 to .9499. Jaden Wiley was paid $66.49 on this day and after the conversion, the amount in Euro was 63.16.

Dynamic import records with more than 100 words:

First a new sheet was added and then the "if" formula was used to import records from the translation data with over 100 words translated. To ensure the condition that only records with more than 100 words can be added to this sheet, data validation of greater than 100 was added to the "word count" column.

Translation summary for records above 100 words

Out of the 1008 total records, only 884 of them were of jobs with more than 100 words translated. There were a total of 12 target languages. They were Italian (it), Danish (da), German (de), Finnish (fi), French (fr), Swedish (sv), Dutch (nl), Spanish (es), Portuguese (pt), Indonesian (id), Japanese (ja), and Thai (th). The top three target languages with the highest number of translation records were Spanish (119), Italian (105), and Swedish (92). Indonesian (19), Dutch (35) and French (43) on the other hand had the least amount of translation jobs.

Total word translated were directly proportional to the number of translation jobs completed. This means that the more translation jobs there were for a specific target language, the more words were translated. Total translation hours on the other hand did not have any type of linear relationship with the number of records. To elaborate, there were a total of 119 translation jobs for the target language of Spanish. It had taken a total of 1745.03 hours to complete these jobs. The Japanese language had only 76 records, but a total of 3701.82 hours were required to complete them. On average, Thai, Japanese, and Indonesian were the top three languages with the highest translation time per record. Italian, Danish, and German on the other hand were the fastest to translate on average per record. A more complete table can be found below.

Language	Number of records	Total words translated	Total translation hours	Average translation hours per record
th	70	23995	3601.52	51.45
ja	76	26268	3701.82	48.71
id	19	6102	526.17	27.69
pt	79	26610	1470.84	18.62
es	119	39161	1745.03	14.66
nl	35	11502	266.97	7.63
sv	92	30675	644.28	7
fr	43	13543	115.21	2.68
fi	69	23245	142.34	2.06
de	86	28875	81.28	0.95
da	91	30180	35.76	0.39
it	105	33704	15.99	0.15

Translation summary for records above 500 words

The number or records with above 500 words were significantly less than records with less than 500 words. There were only a total of 116 such records. Spanish, Italian, and Danish were target languages with the highest number of records. French, Dutch, and Indonesian were target languages with the least number of records. Not surprisingly, Thai, Japanese, and Indonesian were again the target languages that took the longest to translate to per record. As an example, there were only two records with over 500 translated words for Indonesian. The two records had taken on average 41.89 hours to translate per record. French on the other hand had three such records. On average per record however, it had taken only .37 hours. A complete table can be found below.

| Language	Number of records	Total words translated	Total translation hours	Average translation hours per record
th	11	6418	637.27	57.93
ja	12	6865	590.76	49.23
id	2	1193	95.79	41.89
nl	3	1817	54.03	18.01
es	17	10672	256.54	15.09
sv	12	7303	175.25	14.6
pt	11	6933	98.7	8.97
fi	8	4674	30.51	3.81
de	12	7504	16.79	1.4
fr	3	1802	1.11	0.37
da	12	7303	2.84	0.24
it	13	7855	2.6	0.2

Translator summary

There were a total of 24 translators for 12 different target languages. In summary, there were three translators (Ellis, Kacper, Tallulah) for Danish (da), three (Izaak, Kaiden, Robin) for German (de), two (Inaya, Virginia) for Spanish, one (Olivier) for Finnish (fi), four (Brett, Eden, Tamara, Trystan) for French (fr), one (Leonie) for Indonesian, two (Cassandra, Lorraine) for Italian, one (Jaden) for Japanese, two (Madison, Marion) for Dutch, three (Frances, Millie, Penny) for Portuguese, one (Fatma) for Swedish, and one (Tyler) for Thai.

For target languages with multiple translators, some were more efficient than others in regards to translation speed. For example, three translators have worked on jobs for Danish. The average translation time per record for Tallulah Pitts was only .26 hours when compared to 1.9 hours for Ellis O'Doherty. For German, Izaak Daniels was the most inefficient translator because he had the highest average cost per word (.04 Euro) and the longest average translation time per record (2.21 hours). With Spanish, Inaya Lindsay was more efficient with faster transition time and a lower cost per word than Virginia Thompson. Looking at French, Eden Hood was the least efficient translator because it had taken her 91.15 hours to translate one record. For Italian, Cassandra Calhoun was less efficient because she took much longer to translate her work when compared to Lorraine Guerra. Additionally, Cassandra's average cost per word was also higher than that of Larraine. For the Dutch, Marion Fletcher was more efficient than Madison Jarvis. Their average cost per word was the same, however Marion was faster at translating. Finally, for Portuguese, Frances Bloggs was the least efficient in terms of average translation time per record. Penny Moss on the other hand was the most expensive Portuguese translator.

Work day summary

There were a few interesting findings in regards to which day of the week translators preferred to start and finish their work. The most popular days to start a translation job based on the translation data were Monday and Thursday. Wednesday and Tuesday were the least popular days to start work on. Interestingly, no one started their work on the weekend. As for when the translators liked to finish their work, Thursday and Friday were the most popular. Saturday and Sunday were the least popular days to finish work on.
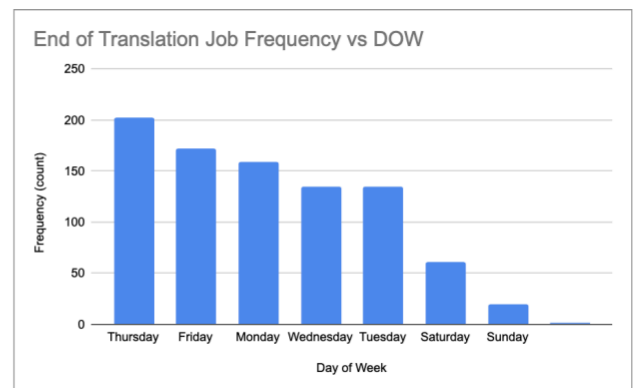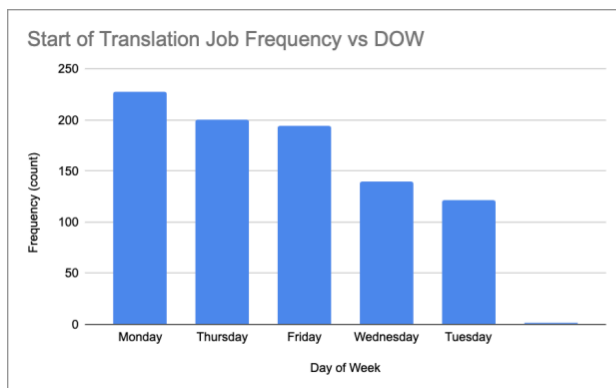
Additional findings:

Out of the 12 target languages, 7 (da, de, es, fr, it, nl, pt) had multiple translators. 5 of them on the other hand only had one translator doing the translations. Assuming all of the target languages are needed, the following recommendations are given based on target languages with multiple translators. The following translators were inefficient with their translation speed when compared to others working in the same target language, namely Ellis O'Doherty, Izaak Daniels, Virginia Thompson, EdenHood, Cassandra Calhoun, Madison Jarvis, and Frances Bloggs. In addition to being slower, some of them even had a higher translation cost per word. For example Izaak Daniels had an average translation time of 2.21 hours per record and an average cost per word of 4 cents Euros. On the other hand, the fastest German translator Kaiden Chamber had an average translation time of .92 hours per record and was paid 3 cents Euros.

Summary of Average Payment per Word And Average Translation Hours per Record
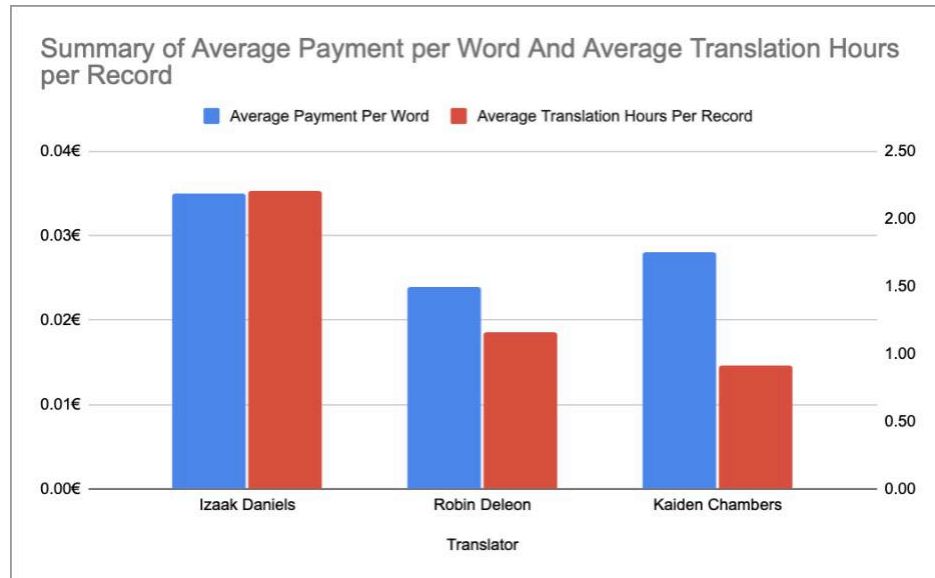
Perhaps a better compensation method would be to adopt a more standardized pay scale based on the target language and translators who are more efficient with translation speed could be paid more.

The Asian languages on average took the longest to translate. However, each had only one translator. Perhaps a more effective strategy would be to add more translators for target languages with high translation time and reduce the number of translators for languages with a low translation time.

Supplementary materials

Perhaps importing the date into a database would be a good step to take. The following are some pseudocode in SQL to import the two data sets.

```sql
'''Creating the translation data table'''

create table translation_data (
    content_id bigserial,
    original_language varchar(2),
    translation_language varchar(2),
    word_count integer,
    status varchar(20),
    created_at timestamp,
    translated_at timestamp,
    translator varchar(100),
    translation_duration_in_hours numeric,
    time_passed_since_translation_days integer,
    constraint content_id_primary_key primary key (content_id)
)
```

```sql
'''Importing translation data into translation data table'''

copy translation_data_table
from 'C:\yourdirectory\translation_data.csv'
with (format csv, header, delimiter ',');
```

```sql
'''Creating the payment data table'''

create table payment_data (
    payment_date date,
    virginia_thompson numeric,
    ...,

)
```

```sql
'''Importing payment data into the payment data table'''

copy payment_data
from 'C:\yourdirectory\payment_data.csv'
with (format csv, header, delimiter ',')
```

The current structure of the payment table may not be totally optimal. A better structure may be of the following. A structure as such would enable joining of the two tables and other relational database features.

```
'''Alternative struture of payment data table'''

create table payment_data (
    payment_date date,
    content_id integer,
    translator varchar(100),
    payment_amount numeric,
    constraint content_id_fk foreign key (content_id)
    references translation_data (content_id)
)
```