EDA

La guerra de Ucrania:

precios, bajas y refugiados

Roger Perelló

ÍNDICE

TEMA

Es una comparativa de 6 *datasets* de tamaños variables sobre diversos aspectos de la guerra de Ucrania que tienen que ver con las pérdidas (sobre todo, las rusas, de las que hay más datos disponibles), los refugiados y los precios en los mercados locales. El nexo de la mayoría son las fechas. Las hipótesis son:

1) Los precios en Ucrania oscilan en función del número de refugiados huidos del país (cuantos más hay, menos demanda) y de la cercanía al frente de los puestos de venta (cuanto más cerca están, más caro es transportar los productos). Asimismo, un peor rendimiento por parte del bando ruso debe de animar a los proveedores extranjeros a vender en el país y aumentar la oferta (con lo cual los productos son más baratos).

A modo de esquema, se puede resumir así:

[Más refugiados, mayor lejanía del frente o más pérdidas rusas = precios más bajos]

2) Con el paso del tiempo, aumentan las pérdidas materiales rusas y se reducen las humanas. Ello es indicativo de una creciente preocupación de la administración rusa por la opinión pública, con lo que, entre otras cosas, se sirve menos de soldados de a pie para minimizar la cuenta de bajas.

A modo de esquema, se puede resumir así:

[Uso mayor de tropas mecanizadas = uso reducido de soldados rasos = más pérdidas materiales y menos muertos]

3) Los ejércitos ruso y ucraniano utilizan mucho material de la Unión Soviética, que, en comparación con el equipamiento moderno, tiende a capturarse con mayor frecuencia.

A modo de esquema, se puede resumir así:

[Uso de material soviético = más destrucciones y más capturas]

OBTENCIÓN DE DATOS

El origen de cada *dataset*, junto con la fuente original y los datos que se le requieren, es el siguiente:

- Acumulado de refugiados ucranianos por fecha y destino (fuente original: API sobre actualizaciones de la guerra de RapidAPI):

https://www.kaggle.com/datasets/anuragbantu/ukraine-invasion-refugee-data-2022

- Precios en cada mercado de Ucrania por producto, fecha, tipo y geolocalización (fuente original: Centre for Humanitarian Data de la Oficina de Naciones Unidas para la Coordinación de Asuntos Humanitarios):

https://data.humdata.org/dataset/wfp-food-prices-for-ukraine

- Pérdidas rusas personales por fecha (fuente original: múltiples, ver en Kaggle; entre ellas, el Ejército y el Ministerio de Defensa del Ucrania):

https://www.kaggle.com/datasets/piterfm/2022-ukraine-russian-war

- Pérdidas rusas materiales, según categorías generales, por fecha (fuente original: la misma, pues sale del mismo Kaggle que el anterior).

- Pérdidas rusas materiales, incluyendo modelo, fabricante, y tipo de baja (captura, destrucción…), sin fecha (fuente original: Oryxspioenkop, web alemana de análisis datos de defensa e investigación sobre la guerra):

https://www.kaggle.com/datasets/piterfm/2022-ukraine-russia-war-equipment-losses-oryx

- Pérdidas ucranianas materiales, incluyendo modelo, fabricante, y tipo de baja (captura, destrucción…), sin fecha (fuente original: la misma, pues sale mismo Kaggle que el anterior).

LIMPIEZA

df_refugees

- Se importa el *dataframe* y no tiene nulos.
- Se crea una función que convierte la columna de fechas ("date") en un *datetime* ordenado (index_by_datetime) y lo lleva al índice, lo cual será útil si hay que unificar tablas.
- Se dejan solo las columnas de país ("country") y acumulado de refugiados ("individuals"). Las demás no interesan para probar la hipótesis.

Salta a la vista que hay varios países que aparecen bastante menos que la mayoría. Sin embargo, como todavía no se ha igualado esta tabla a la otra necesaria para probar la hipótesis (es decir, la de precios), no está claro qué meses harán falta ni qué días de esos meses serán el mejor punto de referencia (por aparecer en la otra). Por tanto, de momento, se dejan todos sin tocarlos.

- Se observa la cabecera del *dataframe* resultante (**Anexo 1.1**).

df_prices

- Se importa el *dataframe* y hay algunos nulos.

Solo con mirar la cabecera de la tabla, y teniendo en cuenta que hay el mismo número de nulos en "admin1", "admin2", "latitude" y "longitude", está claro que corresponden todos al valor de "market" llamado "National Average". Esa media no interesa para el análisis.

- Se inspecciona el índice 0, puramente explicativo, el cual hay que quitar después. Según apunta, las columnas "admin1" y "admin2" son nombres de calles y ciudades, probablemente de donde Naciones Unidas ha recibido los datos... tampoco interesan, y, además, están incompletas. La de "latitude" tampoco hace falta, porque para estimar la cercanía al frente basta el plano horizontal ("longitude", que es la longitud este; cuanto más hacia el este se sitúa una ciudad, más cerca del frente se encuentra). Las columnas desde "priceflag" a "currency" tampoco son útiles, porque todos sus valores son iguales.
- Se eliminan todas las filas donde el mercado es, en realidad, la media nacional ("National Average").
- Se cambian a numéricas las columnas "longitude", "usdprice" y "price".
- Se aplica la función que pone la fecha, en forma de *datetime*, como índice (index_by_datetime).
- Se limita el *dataframe* al período de la guerra con Russia.
- Se comprueba que con todo lo hecho ya no hay nulos.

En este punto toca elegir qué productos (en "commodity") sirven para comparar las oscilaciones de precios en función de las bajas, los refugiados o la lejanía al frente.

- Se empieza por hacer un test ANOVA; aunque es poco probable, si las medias de todos son iguales, cabe coger cualquiera como producto representativo.
- El ANOVA da negativo (pvalor < 0.05), luego las medias no son iguales. Toca ver qué opciones de productos hay.

Tras usar el método .groupby() para las categorías de producto, se nota que, en "non-food", los medicamentos vienen en 3 tipos (antibióticos, antipiréticos y agentes vasodilatadores), y que los tres pueden ser locales o importados; ello ofrece dos dimensiones más para comparar. Además, como en la guerra suele haber más heridos que en períodos de paz, se verá claro con los medicamentos si es cierto que los precios varían de ciudad en ciudad por diferencias en la demanda (que es la causa más probable de que oscilen los precios en un mercado).

- Se eligen los medicamentos para el estudio.

Hay un problema, sin embargo, y es que los antipiréticos importados vienen en una unidad diferente ("1 sachet" vs "10 tablets"). Por tanto, se hace necesario actualizar sus precios para que todos los medicamentos sean comparables.

- Se busca un medicamento en páginas de venta ucranianas que sea de importación, antipirético (como el ibuprofeno) y que se venda tanto en *sachets* (bolsitas) como en *tablets* (comprimidos). Se encuentra este y este.

A 217.90 UAH cada 20 *sachets*, cada *sachet* sale a 10.895 UAH, que son 0.30 dólares. A 109.30 UAH cada 10 *tablets*, cada *tablet* sale a 10.93 UAH, que son, también, 0.30. Por tanto, 1 *sachet* = 1 *tablet*, y hay que multiplicar por 10 todos los valores de "Antypiretic (imported)" para obtener una aproximación razonable que comparar con los demás medicamentos.

- Se hace la corrección y se eliminan las ahora innecesarias columnas de "category" y "unit".
- Se desecha la moneda ucraniana y se deja solo el precio en dólares, ya que es un tipo de cambio más conocido.
- Se observa la cabecera del *dataframe* resultante (**Anexo 1.2**).

df_personnel

- Se importa el *dataframe* y hay algunos nulos.
- Se descartan las columnas “POW” (“prisioneros de guerra”; no es útil para verificar la hipótesis), “personnel*” (solo tiene un valor) y “day” (son los días que lleva la guerra en marcha).
- Se crea una función que permite convertir columnas de valores acumulados en absolutos (decumulate_columns) y se aplica a la lista de bajas “personnel”.
- Se utiliza, de nuevo, la función que pasa la fecha al índice en forma de datetime (index_by_datetime).
- Se borra la primera fecha porque, al haber salido de un acumulado, el valor está inflado por coger datos que no constan en la tabla.
- Se observa la cabecera del *dataframe* resultante (**Anexo 1.3**).

df_equipment

- Se importa el *dataframe* y hay nulos en demasía.
- Se comprueba cómo se distribuyen en porcentajes.
- Se eliminan las columnas “greatest losses direction” y “day”, que no son sobre equipamiento y, por tanto, no valen para corroborar la hipótesis.
- Se investiga en el Kaggle qué hay en las columnas de nulos (**Anexo 1.4**).
- Se convierten a ceros los nulos de las columnas numéricas para poder hacer las sumas correspondientes. Se prevé que algunas de las resultantes tendrán ceros; aquellas en las que sean 0 todos los valores sumados.
- Con la información obtenida, se unen las columnas de temática similar (“fuel tank” y “military auto” pasan a “vehicles and fuel tanks”, mientras que “mobile SRBM system” y “cruise missiles” conforman la flamante “missile systems”).
- Se elimina la columna “special equipment” a pesar de que hay pocos nulos y cabría adjudicarlos a la media o la mediana, pues su contenido no está claro, siquiera, para el autor del *dataset*. Si se diera el caso de que contiene algo fácil de capturar o dañar, como munición o armas ligeras, me desvirtuaría la tabla.
- Se desacumulan las columnas (decumulate_columns) y se lleva la fecha al índice en forma de datetime (index_by_datetime).
- Se añade una columna con la suma de todos los equipamientos.
- Se descarta la primera fecha porque, como viene de un acumulado, se nutre de datos que no aparecen en la tabla.
- Se observa la cabecera del *dataframe* resultante (**Anexo 1.5**).

df_tech

- Se importa el *dataframe* y hay nulos en cantidades industriales.
- Se comprueba el total de nulos. Hay muchos.

Cabe la posibilidad de que muchos de esos nulos sean, de hecho, ceros, porque no ha habido bajas de ese tipo para tal o cual pieza de equipamiento. La forma de saberlo es fijarse en la columna “losses_total”, pues el total de bajas para cada pieza debería ser una suma de todas las subcategorías de capturadas y no capturadas. Si se queda corta, es que faltan datos.

- Se cambian los nulos de las columnas numéricas a ceros.
- Se suman las subcategorías de capturados en una columna que los agrupe, “total captured”, y se hace lo propio con los no capturados.
- Hecha la comprobación, se demuestra que el total de bajas coincide con la suma de las columnas “total captured” y “total not captured”.
- Se observan los no nulos de la columna “sub_model”, que contiene pequeños detalles para unos pocos modelos (“model”) de equipamiento.
- Se fusionan las dos columnas en “model”, pues para resolver la hipótesis, si acaso el modelo llega a ser necesario, bastará con considerar cada combinación de modelo con su submodelo como un modelo propio.
- Se observa la cabecera del *dataframe* resultante (**Anexo 1.6**).

df_uk_tech

- Este *dataframe* es idéntico al anterior, si bien para el bando ucraniano, con lo que el tratamiento es el mismo.
- Por si acaso, se comprueba que la columna “losses_total” equivale a la suma de capturados y no capturados.
- Se observa la cabecera del *dataframe* resultante (**Anexo 1.7**).

HIPÓTESIS

Hipótesis 1

[Más refugiados, mayor lejanía del frente o más pérdidas rusas = precios más bajos]

Hipótesis 1 A:

[Mayor lejanía del frente = precios más bajos]

- Se observa el *dataframe* de precios (df_prices).
- Se crea una función *ad hoc* (create_lmplot_w_regression_line) que dibuja un gráfico de tipo lmplot pero que, dado que será necesario repetirlo, que lo haga con algunos valores por defecto. Así no hará falta escribir lo mismo dos veces. Esta función no va a la carpeta utils porque solo sirve para este caso y no conviene guardarla para futuros análisis.
- Se separa el *dataframe* entre medicamentos locales e importados.
- Se activa la función create_lmplot_w_regression_line para los dos casos (**Anexo 2.1**).

Se nota que los precios de los importados son muy superiores (exceptuando los vasodilatadores). Para los dos casos, además, los precios de los antibióticos y los antipiréticos parecen oscilar más. Por otra parte, algunas de las líneas de regresión bajan un poco al ir acercándose al frente (cuando la longitud este crece), pero es un cambio demasiado pequeño como para considerarse significativo. Así las cosas, una mayor distancia al frente no reduce necesariamente el precio de las medicinas. ¿Ocurre lo mismo con los demás productos?

- Se confirma este hecho con la misma gráfica lmplot que antes, pero aplicada al conjunto de productos (**Anexo 2.2**. Hay que volver a importar el *dataframe* original. Se eliminan los puntos y la leyenda del gráfico para que se vea más claro, además de reducir el tamaño de las líneas); casi todas las líneas de regresión son prácticamente paralelas (si bien hay dos o tres excepciones poco significativas).

Con lo visto, se intuye que un análisis pormenorizado de los medicamentos será más fructífero que uno generalista, que intente abarcar el conjunto de productos. Por esa razón, para el resto de la hipótesis 1 el análisis se focaliza en los antibióticos, los vasodilatadores y los antipiréticos.

- Como ya se necesita la longitud, se comprueba la oscilación de los precios de los medicamentos con la desviación estándar (**Anexo 2.3**).

Es posible que haya una demanda muy superior para los antibióticos y los antipiréticos, la cual hace que los precios sean más altos en general y que oscilen.

- Se comprueba que la oscilación tiene que ver con la guerra con un gráfico que abarca todo el período del *dataframe* original (hay que volver a importarlo), que va más allá de la guerra (**Anexo 2.4**).

Es llamativo que los antipiréticos locales sean tan baratos en comparación a los importados; es posible que se trate de un error humano al montar el *dataframe* original, y que la unidad de medida de los locales sea, como para los importados, "1 sachet", y no "10 tablets".

- Se crea una función que calcula el índice de Gini (calculate_gini) para comprobar si los usuarios notan la oscilación de precios de ciudad en ciudad.

Normalmente este índice se usa para comprobar si un reparto es equitativo; aquí se utiliza para ver si los costes de los medicamentos se reparten de forma “justa” entre los diferentes mercados. Hace falta, para ello, la media (o la mediana) de precios de cada mercado.

- Se observa el resultado (**Anexo 2.5**), que es tan concluyente (muy alejado de 1) que da igual, para el caso, investigar si es mejor servirse de la media o la mediana, así que se deja con la media.

Como el índice de Gini está cerca de 0 para todos los medicamentos, un usuario de a pie no percibe fácilmente que el precio varía de ciudad a ciudad.

Hipótesis 1 B:

[Más pérdidas rusas = precios más bajos]

- Se observa el *dataframe* de bajas de personal (df_personnel), el de pérdidas de equipamiento (df_equipment) y el de precios (df_prices).

Hay que comprobar si los precios oscilan con las bajas rusas personales y materiales. La cuestión es que los precios están calculados mensualmente, mientras que las bajas, día a día. Toca descubrir, para cada mes que aparece en el índice de los precios (df_prices), cuantas bajas ha habido durante el mes previo en las tablas de bajas.

Por ejemplo, si en la tabla de precios está el mes 2022-05-15, el objetivo es sumar las bajas desde 2022-05-14 hasta 2022-04-15... y así para cada mes de la tabla de precios (excepto el primero, ya que se trata de 2022-03-15, y no hay cuenta de bajas antes del 2022-02-25).

- Se crea, con este propósito, una función (date_index_to_monthly).

Con esto ya es factible crear un gráfico lineplot de las bajas mes a mes. Ahora bien, si se quiere una idea general, son necesarios, también, gráficos de precios por mes para cada uno de los tipos de medicamentos (6 en total, contando importados y locales). Como hay que calcular una cifra para cada mes que aglutine todas las ciudades, se hace necesario decidir de nuevo entre la media y la mediana. Esta vez sí que hay que tenerlo en cuenta, ya que no se busca un coeficiente para el cual se prevea un resultado decisivo, sino llevar a cabo una comparativa visual.

- Se elige la mediana, ya que se pretende ver la tendencia con claridad (y que no quede emborronada por algún elemento discordante) y no son tan importantes los valores concretos.
- Se descarta la columna "longitude" de df_prices, que ya no es necesaria.
- Se itera por cada posible medicamento, se saca la mediana de todas ciudades para cada fecha, se borra el índice inicial (que no aparece en la tabla de bajas por razones ya expuestas), y, así, se va creando cada gráfica. Se utiliza una función (apply_condition_groupby_mean_drop) para buena parte del proceso porque será útil más adelante si hay que repetirlo.

Con esto ya es factible crear un gráfico de las bajas mes a mes. Ahora bien, si se quiere una idea general, son necesarios, también, gráficos de precios por mes para cada uno de los tipos de medicamentos (6 en total, contando importados y locales). Como hay que calcular una cifra para cada mes que aglutine todas las ciudades, se hace necesario decidir de nuevo entre la media y la mediana. Esta vez sí que hay que tenerlo en cuenta, ya que no se busca un coeficiente para el cual se prevea un resultado decisivo, sino llevar a cabo una comparativa visual.

Hechas las gráficas (**Anexo 3.1**), se perciben cambios importantes alrededor del sexto mes, con lo que se hará la correlación total entre cada medicamento y las bajas, así como la correlación antes y después de este punto. Antes, sin embargo, hay que elegir entre las bajas de personal y las de equipamiento. Ambas siguen una tendencia similar, pero la de personal parece más pronunciada.

- Se comprueba con el coeficiente de asimetría de Fisher. El de las bajas de personal está mucho más cerca de 1 porque, si bien en ambos casos hay una asimetría por la derecha (se "hunden" por la derecha), en su caso el desequilibrio es mayor. Ello confirma la corazonada que surge al ver la gráfica. Así pues, se decide usar las bajas de personal para el cálculo de la correlación
- Se crea una lista de listas con los índices que hay que eliminar para calcular las 3 diferentes correlaciones (total, antes del mes 6 y después del mes 6).
- Con la misma función de antes (apply_condition_groupby_mean_drop), se crean subconjuntos para cada medicamento con la mediana de los precios de las ciudades para cada fecha, pero, esta vez, se van liquidando los índices que no interesan para cada caso mediante el argumento index_drops de esa función.
- Cada subconjunto se fusiona con un *dataframe* de personal creado con la otra función anterior: date_index_to_monthly. Este *dataframe* es el mismo que el usado para el lineplot de bajas mes a mes, pero con fechas en el índice (es necesario que sea así para que el merge por el índice funcione).
- Se mete en una lista cada vector con los números correspondientes a los meses que interesan para cada correlación, y, luego, todas las listas en una lista más grande.
- Con esa lista de listas ya se puede crear el gráfico de correlación (**Anexo 3.2**).

El resultado es que la correlación cambia totalmente de sentido en ambos casos (antes del mes 6 y después del mes 6), y lo hace de manera contundente (para los 6 medicamentos).

En términos generales, las bajas y los medicamentos están inversamente correlacionados, aunque de un modo poco tajante.

Antes del mes 6, van en la dirección contraria. Sin embargo, a partir del mes 6, van en la misma, y se vuelve muy palpable.

Los medicamentos suben siempre, y las bajas primero bajan y luego suben. Lo que ocurre, con toda probabilidad, es que los medicamentos siempre suben, y las pérdidas primero bajan y luego suben, de ahí que estén correlacionadas.

Así, aunque no sea correcto decir que los precios de los medicamentos vienen dados por las pérdidas de personal, sí se puede asegurar que en el mes 6 las bajas sufren un giro.

La hipótesis 2 estudiará las bajas de personal comparadas con las de equipamiento. Será interesante ver qué ocurre en el mes 6 para que se produzca tal cambio.

Hipótesis 1 C:

[Más refugiados = precios más bajos]

Solo queda comprobar si los refugiados siguen una tendencia similar a lo que ya hemos visto; es decir, si pasa algo con ellos en el mes 6.

- Se observa el *dataframe* de refugiados (df_refugees) y el de precios (df_prices).

La situación es complicada. Se quieren contar los refugiados totales los días 15 de cada mes (los del índice de la tabla mensual de precios, pero sin contar el último, ya que la tabla de refugiados acaba el 2022-09-13). Sin embargo, es de prever que hay países para los que no se hacen las cuentas todos los meses... y mucho menos el día 15 en concreto.

Se hace necesaria una función que, por cada día 15 de esos, mire para qué países hay datos y los sume (recordemos que el *dataframe* de refugiados es un acumulado). Si no encuentra datos para algún país, que mire en la fecha más cercana (el día 16, si acaso existe en el *dataframe*) para obtener una aproximación. Si sigue sin encontrar nada, que mire en la fecha siguiente... y así sucesivamente hasta haber sumado los datos de todos los países para (más o menos) ese día 15. Luego, la función debe hacer lo mismo para el resto de días 15 de cada mes. Al final, debe devolver un *dataframe* con el total de refugiados por fecha.

Hay que señalar que lo que se obtenga será una aproximación, dado que no todos los valores sumados saldrán realmente del 15 de cada mes. Sin embargo, como la fuente es un acumulado con valores bastante grandes, si se procura que no se alejen demasiado de la fecha que les corresponde (no más de 15 días), el resultado es útil al menos para detectar una tendencia general.

- Se crea la función correspondiente (sum_by_duplicated_values_and_datetime). Con cada uso, la función imprime el proceso que va haciendo. Con este método, se puede ir calibrando si, para algún país, se aleja demasiado del día 15 a la hora de hacer su aproximación, y valorar si no queda más remedio que eliminarlo de la tabla. Esta función es compleja, así que conviene ver un ejemplo de su uso (**Anexo 3.3**).

- Se usa la función y solo hay un país que descartar: "Other European countries", que ya se aleja lo indecible de la primera fecha 15 en la primera vuelta, pues no es capaz de rellenar ni un mes (todos sus valores deben ser anteriores al 2022-03-15). Ninguno más, ya que los datos para el resto de países se encuentran antes de llegar al 30 del propio mes, con lo que son una aproximación válida.
- Con la tabla resultante, se desacumula la columna de refugiados (decumulate_columns), se lleva la fecha al índice en forma de datetime (index_by_datetime), se borra el primer mes (ya que no hay con qué desacumularlo) y se modifica el índice para mostrar solo los meses.
- Se hace la gráfica (**Anexo 3.4**). Efectivamente, entre el mes sexto y el séptimo los refugiados se disparan.

Hipótesis 1: conclusiones

A)

- No hay relación entre la cercanía al frente y el precio de los productos. La guerra moderna, con artillería de larga distancia y bombardeos aéreos, y con un frente largo, implica que todas las ciudades sean susceptibles de sufrir problemas puntuales de suministro independientemente de su proximidad al enemigo.
- Los precios de los antibióticos y los antipiréticos, además de ser superiores en términos generales, oscilan mucho. Como el país se encuentra en situación de guerra, es de prever que la demanda de este tipo de medicamentos, muy útiles para una persona herida o con fiebre, se haya disparado. De ahí que oscilen de un modo dramático comparados con los agentes vasodilatadores, que sirven para tratar la tensión arterial, un problema de salud cuyos factores de riesgo no se ven agravados en un conflicto bélico.
- Para muestra, Mayo Clinic ofrece un listado de factores de riesgo para una tensión arterial alta. Cuestiones como la edad, el sobrepeso o el consumo de tabaco no se agravan en un conflicto como la posibilidad de sufrir fiebre o infecciones.
- Es poco probable que las diferencias en el precio de los medicamentos de ciudad a ciudad, por sí solas, supongan un problema para los consumidores, ya que el índice de Gini es bajo para todas; el único problema es que la demanda suba tanto que estos lleguen a escasear (o que la oferta se reduzca).

B)

- Solo en términos de correlación, el precio de la medicina sube, mientras que las pérdidas primero bajan y luego suben.
- Es a partir del mes 6 que las bajas rusas vuelven a subir.

C)

- Los refugiados mensuales se disparan entre el mes sexto y el séptimo, cuando las bajas rusas se encuentran en un punto bajo (es posible que la guerra se recrudezca).
- Se ha desmentido de forma prematura la hipótesis 2 con el lineplot de bajas (las bajas personales y materiales rusas no siguen tendencias diametralmente opuestas). Sin embargo, sigue adelante la investigación para encontrar pistas sobre lo que pasa alrededor del sexto mes.

Hipótesis 2

[Uso mayor de tropas mecanizadas = uso reducido de soldados rasos = más pérdidas materiales y menos muertos]

- Se observa el *dataframe* de pérdidas de equipamiento (df_equipment).
- Se empieza con una gráfica boxplot (**Anexo 4.1**), según la cual 1) hay diferencias importantes en los números de bajas; 2) los *outliers* de las bajas totales ("total losses"), con toda probabilidad, vienen sobre todo de las bajas de APC - que, además, son mayoritarias- y de los "vehicles and fuel tanks"; y 3) no hay *outliers* en el límite inferior de las bajas totales, lo que significa que no cabe preocuparse por números exageradamente bajos si, por alguna razón, se acaba mirando por debajo de la mediana.
- Se percibe, también, un error relativo a las bajas de APC. Están por debajo de 0, y no puede haber bajas negativas.
- Una vez localizado el error, se pregunta al creador del *dataset*. Parece que es una mera corrección de los datos acumulados.
- Se ignora el fallo porque es algo muy pequeño y puntual como para afectar a la visión conjunta.
- Se comprueba hasta qué punto las bajas por “APC” y las de “vehicles and fuel tanks” son mayoritarias con un gráfico de barras (**Anexo 4.2**).

Con lo visto en las hipótesis previas, se deduce que a partir del mes 6 hay una caída en las bajas.

- Se comprueba con dos gráficos: uno con las bajas de APC + “vehicles and fuel tanks” y las pérdidas totales, y otro con el resto de tipos de bajas y las pérdidas totales (**Anexo 4.3**).

En efecto, se detecta una ligera bajada que va del mes 6 al 9. Cuando vuelven a subir las bajas totales, las de APC y “vehicles and fuel tanks” ya no son tan fuertes.

Para comprobar esa bajada, es necesaria la probabilidad de que un conjunto de bajas diarias situado entre el día uno del mes sexto (incluido por estar a la izquierda, como es habitual al calcular intervalos), y el del noveno (no incluido) esté por debajo de la mediana.

Se elije la mediana por dos razones: 1) hay *outliers*, de ahí que se prefiera antes que la media, y 2) no hay *outliers* por debajo de la mediana, como demuestra el boxplot previo.

- Se calcula la probabilidad condicionada tal que así:

$P(Mes \cap Med)$
$= \ probabilidad\ de\ que\ un\ conjunto\ al\ azar\ de\ bajas\ diarias\ esté\ entre\ el\ mes\ sexto\ y\ el\ noveno\ y\ por\ debajo\ de\ la\ mediana$

$P(Mes) = probabilidad\ de\ que\ un\ conjunto\ de\ bajas\ diarias\ esté\ entre\ el\ mes\ sexto\ y\ noveno$

$P(Med/Mes)$
$= probabilidad\ de\ que\ un\ conjunto\ entre\ el\ mes\ sexto\ y\ el\ noveno\ esté\ por\ debajo\ de\ la\ mediana$

$$P(Med/Mes) = \frac{P(Mes \cap Med)}{P(Mes)}$$

$$P(Med/Mes) = 89.130\%$$

Queda demostrado matemáticamente que hay una bajada entre los meses 6 y 9.

Hipótesis 2: conclusiones

- La mayoría de bajas son de APC. Esas son las siglas de Armored Personnel Carrier, lo cual implica que hay bajas de personal simultáneamente, ya que se destruyen vehículos dedicados al transporte de personal, si bien con capacidades defensivas
- La columna "vehicles and fuel tanks", que también es mayoritaria, incluye, a todas luces, los vehículos sin capacidad de autodefensa de transporte de personal junto con los de transporte de combustible. Así pues, esta categoría y la de APC incorporan fundamentalmente vehículos de transporte de infantería
- La mayor parte de bajas diarias para el período del sexto al noveno mes están por debajo de la mediana
- Como hemos visto en el primer gráfico de "Hipótesis 1 B", las bajas de personal para este período (aproximadamente) también se encuentran en un punto bajo
- Se deduce, pues, que el hecho de que haya menos bajas de personal (y no solo de equipamiento) tiene que ver con que se han destruido menos APC y/o vehículos de transporte convencionales en este período, pues la destrucción de cualquiera de los dos tipos de vehículo supone, irremediablemente, la muerte de sus tripulantes y pasajeros
- En el mes 6 también suben mucho los refugiados. Para este período, es probable que Rusia se esté apoyando más en bombardeos y artillería que en la conquista de territorio con tropas terrestres. Eso explicaría tanto el aumento de refugiados (se puede vivir en una ciudad ocupada, pero no devastada) como la menor pérdida de convoyes de transporte -y, por ende, de personal en general-. El hecho de que, cuando vuelven a subir las bajas de personal (visto en la hipótesis 1 B) y de equipamiento totales (visto en el último gráfico), alrededor del mes 9, las bajas de vehículos de transporte ya no suban tanto, da fuerza a esta proposición.

Hipótesis 3

[Uso de material soviético = más destrucciones y más capturas]

- Se observan los *dataframes* de pérdidas materiales rusas sin fecha (df_tech) y de pérdidas ucranianas (df_uk_tech), con el foco en el ruso.
- Para el *dataframe* ruso, se hace una matriz de correlación porque tiene muchas variables numéricas que tienen que ver las unas con las otras (**Anexo 5.1**).
- Se detecta una correlación anómala entre el equipamiento abandonado (que cuenta como no capturado) y el equipamiento capturado total.
- Se comprueba si sus medias son iguales mediante un ftest + ttest que sirve como indicio para saber si vale la pena estudiar esa correlación. No lo son.
- Se hace lo propio con la pareja de equipamiento capturado (“captured”, una subcategoría) y equipamiento capturado total (“total captured”, categoría general), y con “destroyed” y “total not captured”.

Las medias son iguales; si fuera necesario, sería razonable, pues, referirse al total de capturados y no capturados simplemente como "capturados" y "destruidos" respectivamente.

- Se crean dos tablas para el *dataframe* ruso: una de equipamiento soviético y otra de no soviético, y se representan gráficamente (**Anexo 5.2**).

Los vehículos de transporte de infantería, tanto los que tienen capacidad ofensiva ("infantry fighting vehicles", llamados APC en la hipótesis anterior) como los inofensivos ("trucks, vehicles and jeeps") son mayoría para los conjuntos soviético y no soviético respectivamente.

Existe una categoría a parte de "APC" (Armoured Personnel Carrier) en estos gráficos, también, pero eso y "infantry fighting vehicles" (IFV) son prácticamente lo mismo, con pequeñas (y muy discutidas) diferencias (el uno estaría más orientado al transporte, y, el otro, al combate). Si en la tabla de equipamiento de la hipótesis anterior no se nombra otra tipología de blindado de transporte que el APC, es porque ya aglutina todos los vehículos de transporte con capacidad de combate (APC e IFV), no porque no existan. De lo contrario, las tablas estudiadas (df_equipment y df_tech/df_uk_tech) se estarían contradiciendo.

Por norma general, la mayoría de equipamiento de un ejército moderno lo forman los vehículos terrestres, y que el grueso de sus vehículos son, en efecto, automóviles blindados; también para el caso de Ucrania y Rusia (**Anexo 5.3**). El enlace solo indica las capacidades militares de ambas naciones; por razones obvias, no se dará a conocer el número real de tropas desplegadas hasta que pase la guerra.

- Se crea, para los *dataframes* ruso y ucraniano, mediante un par de funciones *ad hoc* (que no van a utils), una columna que será la proporción de capturas respecto al total. El objetivo es comprobar si la proporción en capturas es superior para el equipamiento soviético respecto al equipamiento no soviético.
- Se preparan dos gráficos (**Anexo 5.4**) que no son concluyentes.

Es necesario comprobar por otras vías si la media de los soviéticos o la de no soviéticos es superior, tanto para Rusia como para Ucrania, además de cuál de los dos tiene una mayor media de capturas.

- Son pocos valores para cada caso, así que, una vez comprobado que las varianzas son iguales mediante ftests, se hacen ttests.

Todos los pvalores son superiores a 0.05, y, por tanto, iguales. Ni se capturan más (ni menos) soviéticos en ningún caso, ni hay diferencia entre lo que captura Rusia y lo que captura Ucrania.

Hipótesis 3: conclusiones

- La mayoría de pérdidas del bando ruso son vehículos de transporte de infantería armados (soviéticos) y vehículos de transporte no armados (no soviéticos).
- Ambas categorías corresponden a APC y "vehicles and fuel tanks", respectivamente, para el caso de la hipótesis previa.
- Los ttests demuestran que no hay diferencia entre el número de capturas para el equipamiento soviético y no soviético. Tampoco la hay en las capturas totales de Rusia y las Ucrania.

CONCLUSIONES FINALES

- A la hora de enviar medicina a las ciudades, no hace falta aplicar reducciones de precio en función de la cercanía al frente.
- Hay que pedir a los aliados que envíen más antibióticos y antipiréticos con tal de bajar los precios y frenar la oscilación.
- Hay que fijarse en los cambios de estrategia rusa: en el momento en que se sirvan menos de infantería (probablemente porque pasen a destruir las ciudades en lugar de tomarlas), los países colindantes deben prepararse para la posibilidad de recibir más refugiados.
- La mejor manera de incrementar las bajas de personal de Rusia es atacar a sus vehículos de transporte de infantería, en lugar de emboscar soldados a pie. Del conjunto de automóviles, lo más práctico es atacar camiones y otros no blindados. Si hay que acometer a vehículos de transporte con capacidad de autodefensa y es factible hacer la distinción, es preferible que sean APC (Armoured Personnel Carrier) antes que IFV (Infantry Fighting Vehicle). Si se atacan APC/IFV, mejor que sean soviéticos; si se atacan vehículos no blindados, que sean no soviéticos.
- No hay que discriminar a la hora de dar uso a equipamiento soviético o no soviético; al menos, si solo se tiene en cuenta la posibilidad de que sea destruido o capturado.

ANEXOS

Anexo 1.1

date	country	individuals
2022-03-01	Belarus	341
2022-03-01	Poland	453982
2022-03-01	Hungary	116348
2022-03-01	Republic of Moldova	79315
2022-03-01	Russian Federation	42900

Anexo 1.2

date	market	longitude	commodity	usdprice
2022-03-15	Rivne	26.251617	Antibiotics (imported)	4.8148
2022-03-15	Rivne	26.251617	Antibiotics (local)	1.0399
2022-03-15	Rivne	26.251617	Antipyretic (local)	0.4625
2022-03-15	Rivne	26.251617	Vasodilating agents (local)	0.3667
2022-03-15	Rivne	26.251617	Vasodilating agents (imported)	1.5907

Anexo 1.3

date	personnel
2022-02-26	1500
2022-02-27	200
2022-02-28	800
2022-03-01	410
2022-03-02	130

Anexo 1.4

- Military Auto - has not been tracked since 2022-05-01; joined with `Fuel Tank` into `Vehicles and Fuel Tanks`
- Fuel Tank - has not been tracked since 2022-05-01; joined with `Military Auto` into `Vehicles and Fuel Tanks`
- Anti-aircraft warfare
- Drone - UAV+RPA
- Naval Ship - Warships, Boats
- Anti-aircraft Warfare
- Mobile SRBM System - has not been tracked since 2022-05-01; joined into `Cruise Missiles`
- Vehicles and Fuel Tanks - appear since 2022-05-01 as a sum of `Fuel Tank` and `Military Auto`
- Cruise Missiles - appear since 2022-05-01

Anexo 1.5

date	aircraft	helicopter	tank	APC	field artillery	MRL	drone	naval ship	anti-aircraft warfare	vehicles and fuel tanks	missile systems	total losses
2022-02-26	17	19	66	190	0	0	2	0	0	30.0	0.0	324.0
2022-02-27	0	0	4	0	1	0	0	0	0	0.0	0.0	5.0
2022-02-28	2	3	0	110	24	17	1	0	5	161.0	0.0	323.0
2022-03-01	0	0	48	30	3	3	0	0	2	14.0	0.0	100.0
2022-03-02	1	2	13	16	8	16	0	0	2	50.0	0.0	108.0

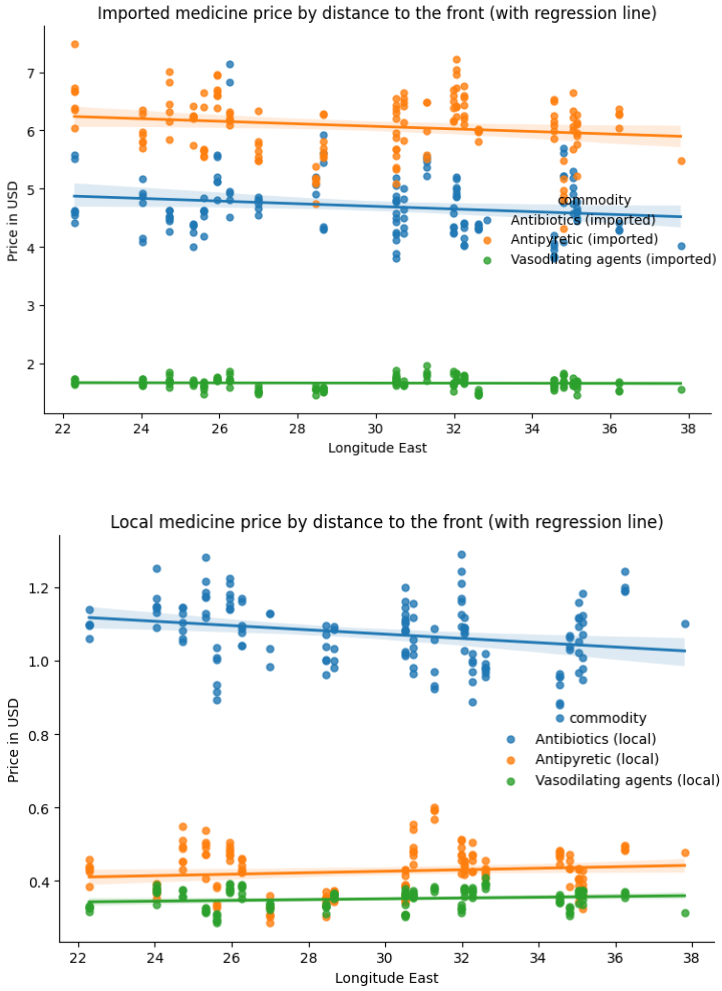
Anexo 1.6

	equipment	model	manufacturer	losses_total	abandoned	abandoned and destroyed	captured	captured and destroyed	captured and stripped	damaged	damaged and abandoned	damaged and captured	damaged beyond economical repair	damaged by Bayraktar TB2	destroyed	destroyed by Bayraktar TB2	destroyed by Bayraktar TB2 and Harpoon AShM	sunk	total captured	total not captured
0	Tanks	T-62M	the Soviet Union	20	1.0	0.0	14.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	3.0	0.0	0.0	0.0	16.0	4.0
1	Tanks	T-62MV	the Soviet Union	3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	2.0	1.0
2	Tanks	T-64A	the Soviet Union	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	2.0
3	Tanks	T-64BV	the Soviet Union	39	2.0	0.0	4.0	0.0	0.0	2.0	0.0	1.0	0.0	0.0	30.0	0.0	0.0	0.0	5.0	34.0
4	Tanks	T-72A	the Soviet Union	33	1.0	0.0	15.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	16.0	0.0	0.0	0.0	15.0	18.0

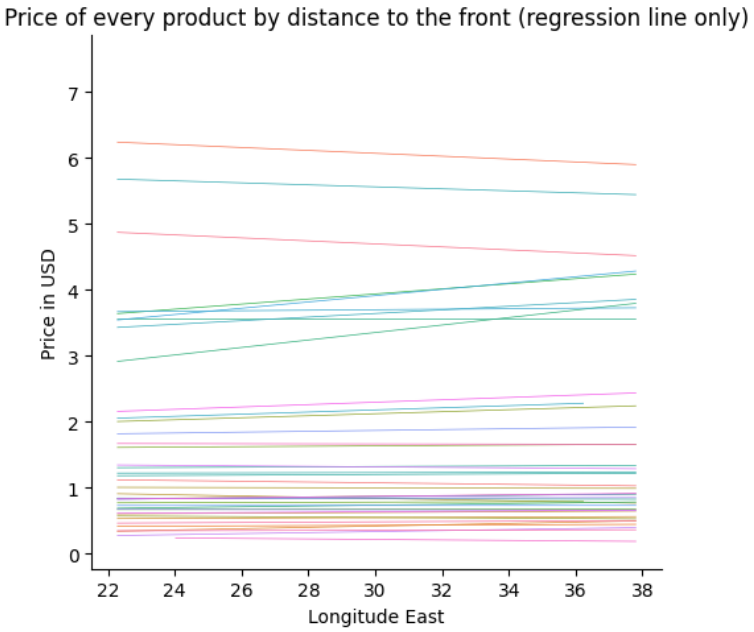
Anexo 1.7

	equipment	model	manufacturer	losses_total	abandoned	abandoned and destroyed	captured	captured and destroyed	damaged	damaged and abandoned	...	damaged by Orion and captured	destroyed	destroyed by Forpost-R	destroyed by Orion	destroyed by loitering munition	scuttled to prevent capture by Russia	sunk	sunk but raised by Russia	total captured	total not captured
0	Tanks	T-64A	the Soviet Union	1	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
1	Tanks	T-64B	the Soviet Union	1	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
2	Tanks	T-64BV	the Soviet Union	123	3.0	0.0	41.0	8.0	3.0	1.0	...	0.0	63.0	0.0	0.0	0.0	0.0	0.0	0.0	53.0	70.0
3	Tanks	T-64BV Zr. 2017	Ukraine	49	3.0	0.0	27.0	0.0	1.0	1.0	...	0.0	17.0	0.0	0.0	0.0	0.0	0.0	0.0	27.0	22.0
4	Tanks	T-64B1M	Ukraine	4	0.0	0.0	4.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0

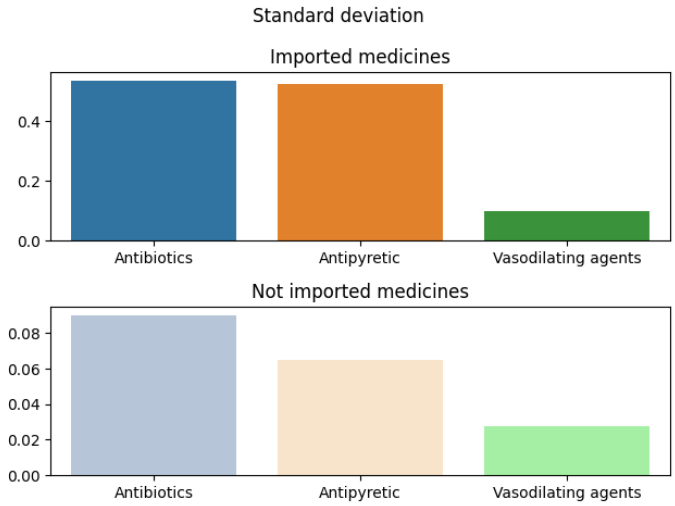
Anexo 2.1



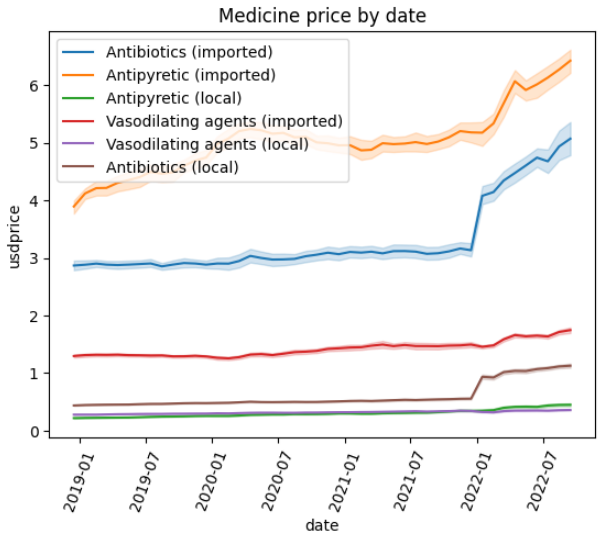
Anexo 2.2



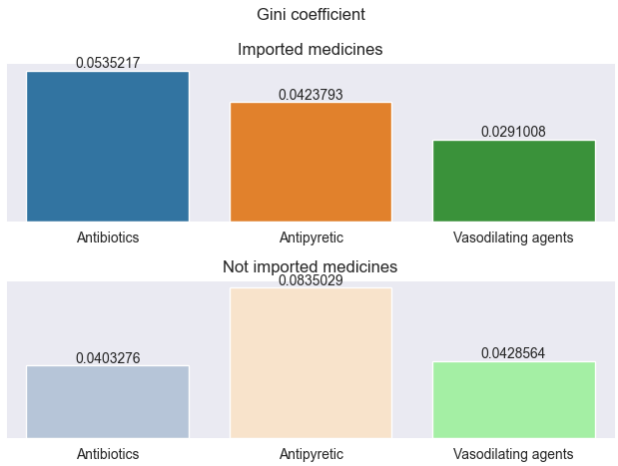
Anexo 2.3



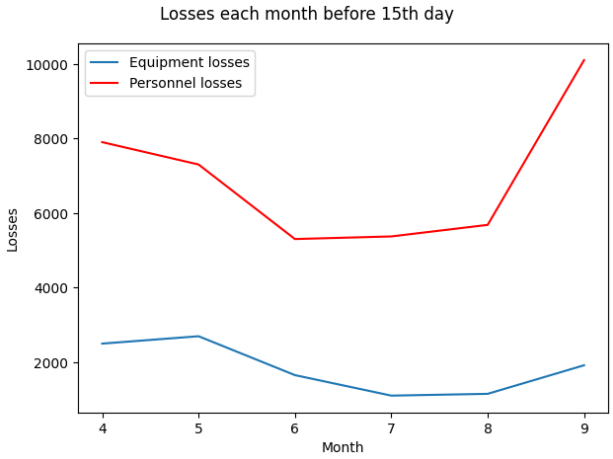
Anexo 2.4

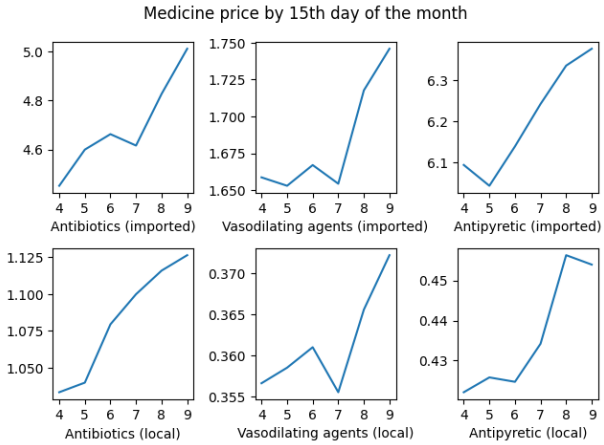


Anexo 2.5

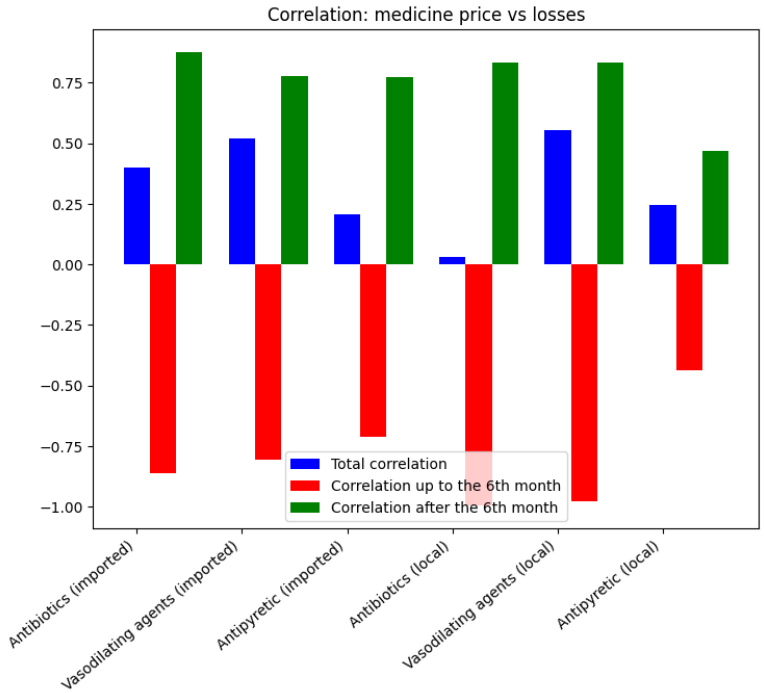


Anexo 3.1





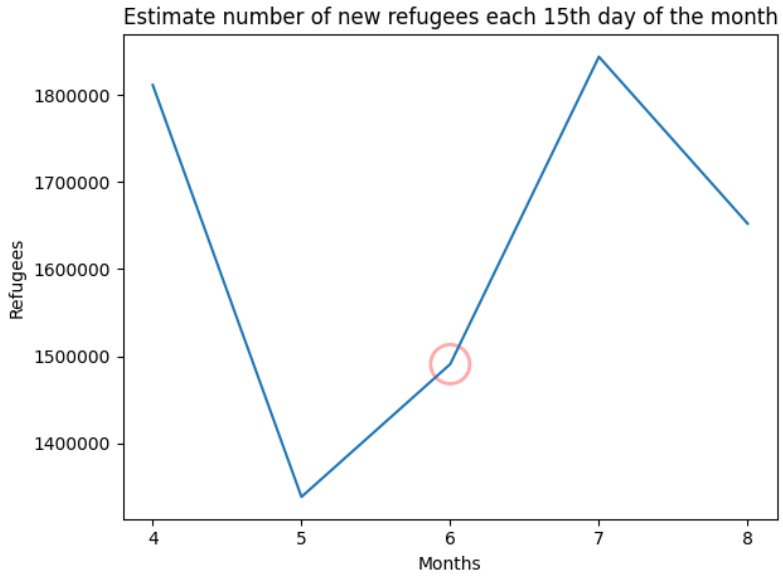
Anexo 3.2



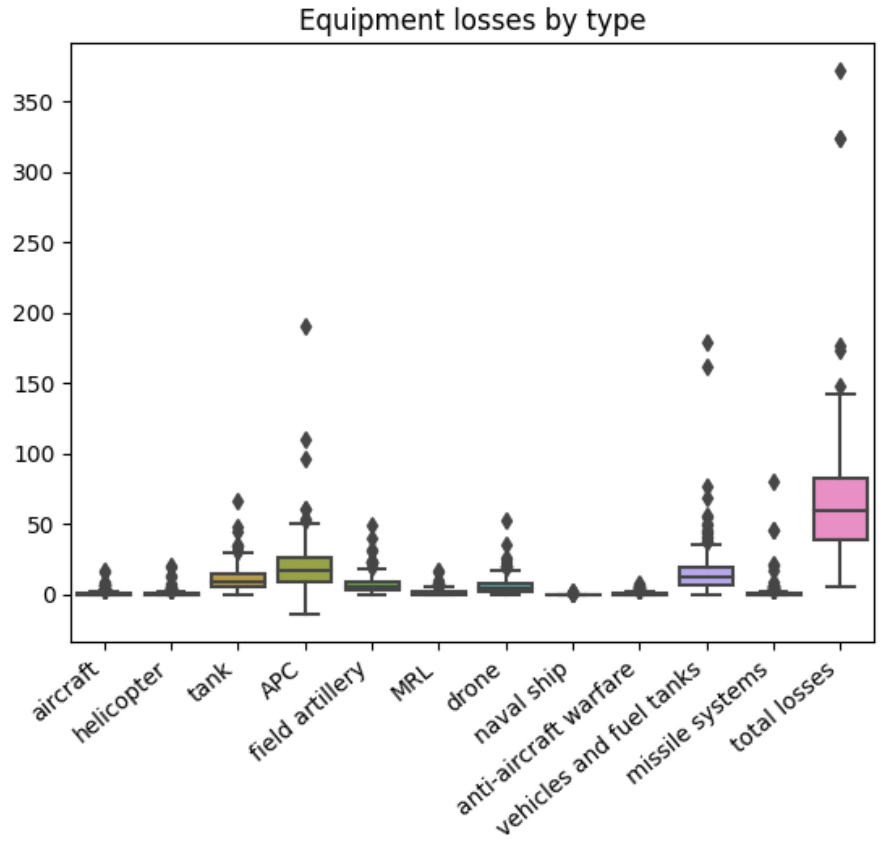
Anexo 3.3

```
Now checking for 2022-03-15 00:00:00
2022-03-15 00:00:00 has been filled


Now checking for 2022-04-15 00:00:00
Missing: ['Russian Federation', 'Belarus']
Add 1 day/s. Current date: 2022-04-16 00:00:00
Missing: ['Russian Federation', 'Belarus']
Add 1 day/s. Current date: 2022-04-17 00:00:00
Missing: ['Russian Federation', 'Belarus']
Russian Federation found at 2022-04-17 00:00:00
Belarus found at 2022-04-17 00:00:00
2022-04-15 00:00:00 has been filled


Now checking for 2022-05-15 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-16 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-17 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-18 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-19 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-20 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-21 00:00:00
2022-05-21 00:00:00 does not exist in the dataframe
Add 1 day/s. Current date: 2022-05-22 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-23 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-24 00:00:00
Missing: ['Belarus']
Add 1 day/s. Current date: 2022-05-25 00:00:00
Missing: ['Belarus']
Belarus found at 2022-05-25 00:00:00
2022-05-15 00:00:00 has been filled


Now checking for 2022-06-15 00:00:00
Missing: ['Hungary', 'Slovakia', 'Romania', 'Republic of Moldova', 'Russian Federation']
Add 1 day/s. Current date: 2022-06-16 00:00:00
Missing: ['Hungary', 'Slovakia', 'Romania', 'Republic of Moldova', 'Russian Federation']
Hungary found at 2022-06-16 00:00:00
Slovakia found at 2022-06-16 00:00:00
Romania found at 2022-06-16 00:00:00
Republic of Moldova found at 2022-06-16 00:00:00
Russian Federation found at 2022-06-16 00:00:00
2022-06-15 00:00:00 has been filled


Now checking for 2022-07-15 00:00:00
2022-07-15 00:00:00 does not exist in the dataframe
Add 1 day/s. Current date: 2022-07-16 00:00:00
```

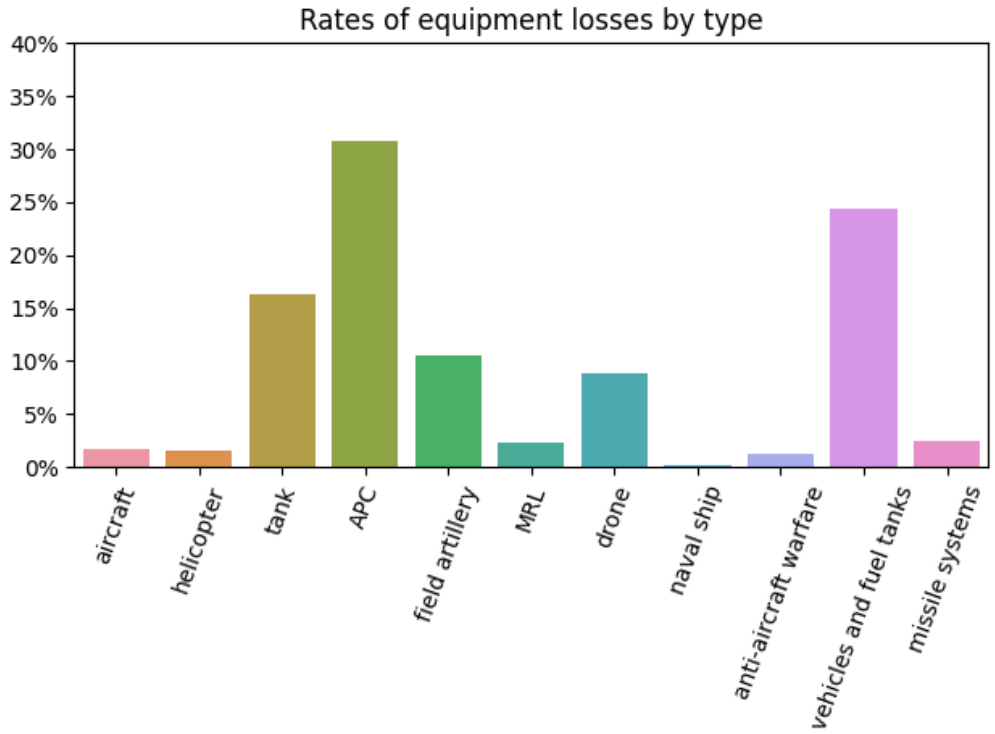
Anexo 3.4



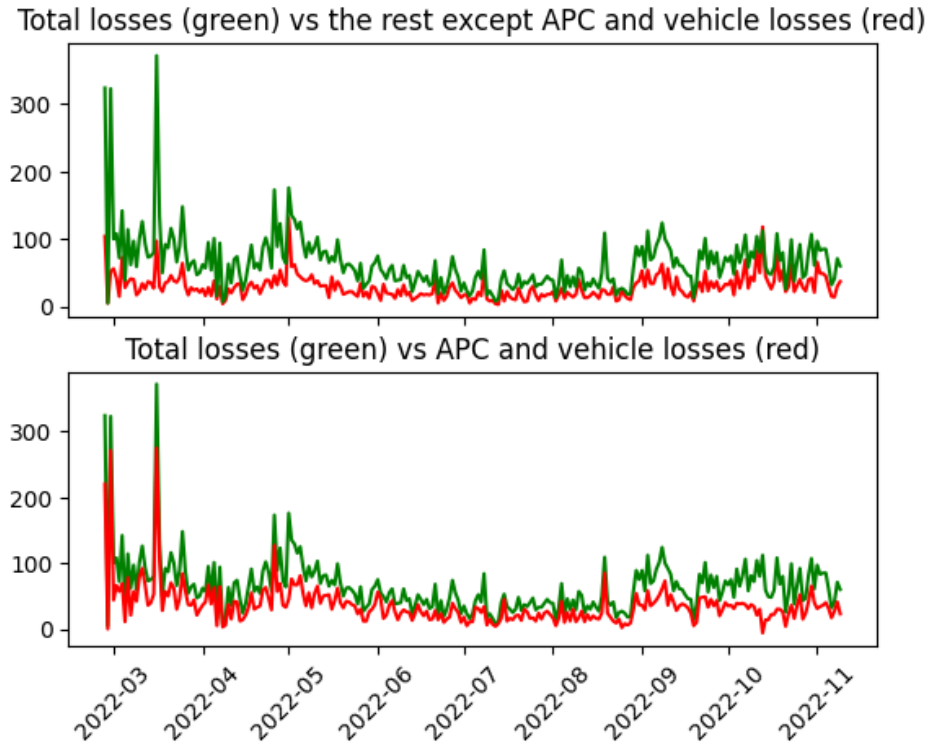
Anexo 4.1



Anexo 4.2

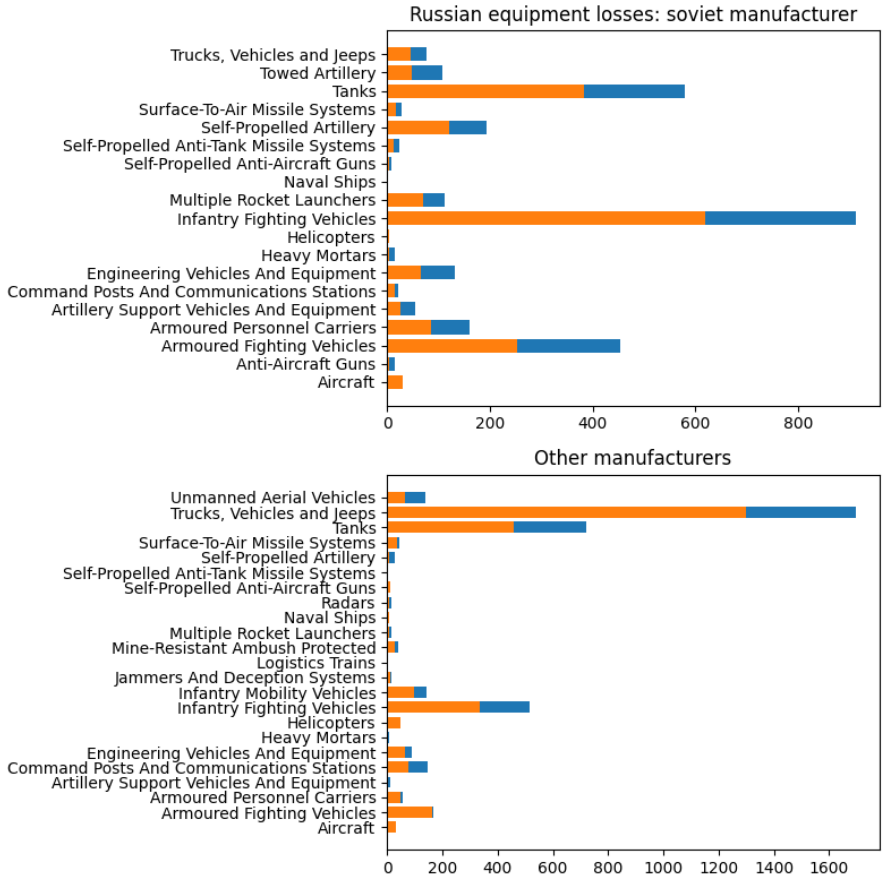


Anexo 4.3



Anexo 5.1

Anexo 5.2



Anexo 5.3

Indicator	Russia	Ukraine
Special-mission aircraft	132	5
Ground combat vehicles	-	-
Armored vehicles	30,122	12,303
Main battle tanks	12,420	2,596
Tower artillery	7,571	2,040
Self-propelled artillery	6,574	1,067
Mobile rocket projectors	3,391	490
Naval forces	-	-
Total military ships	605	38

PDF XLS PNG PPT

Sources
- Show sources information
- Show publisher information
- Use Ask Statista Research Service

Release date
February 2022

Region
Russia, Ukraine

Anexo 5.4