

TRANSPARENCIA DE DATOS SOBRE EL COVID- 19



Grado Ciencia de Datos

Carlos Gallego Andreu

Alejandro Losa Brito

Héctor Martínez Cabanes

Daniel Oliver Belando

Daniel Romero Alvarado

ÍNDICE:

1. [Introducción](#)
 - 1.1 [Palabras clave](#)
 - 1.2 [Términos que conocer para una mejor comprensión del trabajo](#)
2. [Objetivos](#)
3. [Hipótesis de la investigación](#)
4. [Variables](#)
5. [Gestión de los datos recogidos](#)
6. [Ficha técnica](#)
7. [Selección de la muestra y tratamiento de los datos](#)
8. [Colectivos con los que se compartirán los resultados](#)
9. [Análisis DAFO](#)
10. [Análisis](#)
 - 10.1 [Univariante](#)
 - 10.2 [Bivariante](#)
11. [Inferencia sobre la población](#)
12. [Conclusiones objetivas de la transparencia de datos sobre el COVID-19](#)
 - 12.1 [Gobierno central](#)
 - 12.2 [Gobierno autonómico](#)
13. [Conclusión](#)
14. [Bibliografía](#)

Introducción

Debido a la crisis generada por el COVID-19, se ha planteado la pregunta sobre el grado de transparencia del gobierno central con los datos sobre la pandemia. No obstante, los datos son proporcionados por las Comunidades Autónomas, de manera que su grado de transparencia adquiere importancia para la sociedad. A partir, de esta pregunta inicial, hemos realizado un formulario para recoger información de la opinión pública de la ciudadanía. De esta forma, podremos comparar lo que percibe la sociedad con la realidad presentada. Además, a partir de la muestra recogida en la encuesta se pretende extrapolar conclusiones para la población española.

Por otro lado, la intención del proyecto es agregar valor de interés común como puede ser la opinión de la población sobre la gestión del gobierno central. También sabremos qué información resulta más importante conocer respecto al nivel de gravedad de las personas con COVID-19 y los medios más utilizados para informarse. Además, se identificará qué porcentaje poblacional ha sido víctima de adquirir información no confirmada por las fuentes oficiales, es decir, mediremos el porcentaje de personas que han confiado en la información que contenía un bulo.

A todo lo mencionado anteriormente es esencial sumar que tendremos la capacidad de saber qué porcentaje de las personas no sabe interpretar un gráfico, en concreto un histograma, que es el más utilizado para representar la evolución de la pandemia.

Palabras clave:

- COVID-19
- ESPAÑA
- GOBIERNO
- PANDEMIA
- DATOS
- TRANSPARENCIA DE DATOS

Términos que conocer para una mejor comprensión del trabajo

Datos abiertos: Los datos abiertos son datos que están disponibles para ser reutilizados y redistribuidos por cualquier persona. Por tanto, estos paquetes de datos deben ser fácilmente accesibles y deben estar disponibles para cualquier persona o colectivo, sin haber ninguna discriminación en este aspecto.

Además, los datos abiertos son modificables e integrables en otros conjuntos de datos, y pueden ser usados tanto de forma comercial, como en escuelas o universidades.

Agenda 2030: La Agenda 2030 o los Objetivos de Desarrollo Sostenible (ODS), es un proyecto impulsado por la ONU y respaldado por más de 180 países que tiene como objetivo asegurar y mejorar el progreso social y económico de una manera sostenible en todo el mundo. Este proyecto recibe su nombre ya que se puso como fecha para evaluar los resultados el año 2030

La Agenda 2030 está compuesta por 17 Objetivos de Desarrollo Sostenible, muy generales, los cuales a su vez se dividen en 230 indicadores globales. Estos indicadores consisten en objetivos mucho más concretos para alcanzar los 17 principales.

Este trabajo se puede catalogar dentro de los siguientes indicadores de la Agenda 2030:

OBJETIVO 16 Promover sociedades pacíficas e inclusivas para el desarrollo sostenible, facilitar el acceso a la justicia para todos y construir a todos los niveles instituciones eficaces e inclusivas que rindan cuentas.

Concretamente:

- 16.6 Instituciones eficaces y transparentes

Hemos decidido calcular un indicador a partir de nuestros datos, ya que en este apartado la opinión de la población tiene un gran peso, calcularemos la proporción de individuos que creen que los datos son transparentes, para el gobierno central, autonómico y local.

Para ello dividiremos los que hayan puntuado 3, 4 o 5 entre todos los que han contestado (la puntuación está entre 1 y 5) excluyendo los que han contestado “no lo sé” o nulo, más adelante con el estudio univariante se puede entender mejor de donde salen estos datos:

1. ¿Ha sido el gobierno central transparente?

Positivo: 235, **Total:** 577, Proporción: **40.72%**, Se podría decir que no se cumple el indicador. Debe superar el 50%.

2. ¿Ha sido el gobierno autonómico transparente?

Positivo: 268, **Total:** 539, Proporción: **49,72%**, Se podría decir que no se cumple el indicador, pero se aproxima mucho. Debe superar el 50%.

3. ¿Ha sido el gobierno local transparente?

A favor: 232, **Total:** 492, Proporción: **47.15%**, Sigue sin superar la mitad de las respuestas. No cumple el indicador. Debe superar el 50%.

- 4.6 De aquí a 2030, asegurar que todos los jóvenes y una proporción considerable de los adultos, tanto hombres como mujeres, estén alfabetizados y tengan nociones elementales de aritmética

Este indicador podemos calcularlo a partir de los niveles de educación que han contestado al cuestionario, así como viendo si han acertado la pregunta del gráfico, donde había 3 erróneas y una correcta.

Codificando los valores del año de nacimiento en 3 grupos, y mediante una tabulación cruzada con las variables de gráfico y estudio, podemos calcular el porcentaje que han contestado bien. Si más de $\frac{2}{3}$ han acertado, se cumpliría el indicador.

Mediante un sistema de puntos podemos valorar sus estudios

Primaria/EGB: 1 pt, ESO: 3 pt, BACH/FP: 5 pt, GRADO/OPOSICIONES/MASTER: 7 pt,
Una media mayor que 3.5 indicaría que se cumple el indicador.

1. Personas mayores (1940-1971)

- a. ¿Han contestado bien al gráfico?:
Bien: 112, Total: 165, Proporción: 67.8%, Se cumple el indicador, aunque es el grupo con peor ratio.
- b. ¿Cómo son sus estudios?:
Total 951 pt, /166 personas = 5.72. De media nivel de bachillerato, se cumple el indicador

2. Personas adultas (1972-1992)

- a. ¿Han contestado bien al gráfico?:
Bien: 119, Total: 155, Proporción: 76.77%, Se cumple el indicador
- b. ¿Cómo son sus estudios?:
Total 853 pt, /153 personas = 5.57. De media nivel de bachillerato, se cumple el indicador

3. Jóvenes (1992-2006)

- a. ¿Han contestado bien al gráfico?:
Bien: 210, Total: 278, Proporción: 75.53%, Se cumple el indicador
- b. ¿Cómo son sus estudios?:
Total 1558 pt, /282 personas = 5.52. De media nivel de bachillerato, se cumple el indicador.

Objetivos del trabajo

La misión del trabajo es saber si, en términos generales, la población española está bien informada sobre el Covid-19, esta misión se puede desglosar en los siguientes objetivos principales:

- Recopilación de una muestra de la población española a través de una encuesta anónima.
- Realizar un análisis de los datos recogidos en la encuesta.
- Realizar inferencia para extraer conclusiones respecto de la población española.
- Comparar las conclusiones extraídas con nuestras hipótesis iniciales.

Hipótesis de la investigación

Hoy en día las redes sociales son un medio de comunicación muy efectivo, pero también son un canal de difusión de información falsa, mal contrastada o incompleta. Nuestra hipótesis de partida en este trabajo se basa en que una gran parte de la población no contrasta la información a la que accede con fuentes oficiales o contrastadas sobre el tema que se está informando. Por otra parte, también pensamos que muchos de los gráficos que se publican en periódicos, televisiones, etc no son bien interpretados lo que también da lugar a conclusiones erróneas.

Después de realizar todos los análisis de inferencia, compararemos estas hipótesis iniciales con los resultados obtenidos a través de intervalos de confianza para extraer las conclusiones sobre el trabajo.

Variables

1. Indique los métodos de propagación del COVID-19 según las fuentes oficiales.
2. Indique su año de nacimiento (4 cifras).
3. Indique su sexo
4. Marque la opción que mejor describa su situación
5. Indique la Comunidad Autónoma a la que pertenece
6. Nivel máximo de estudios
7. Indique su código postal
8. En una escala del 1 al 5 valore la rapidez con la que el gobierno está tomando medidas contra la pandemia
9. En una escala del 1 al 5 valore el nivel de transparencia de los datos del COVID-19 del gobierno central, autonómico y local
10. ¿Cuál es la información acerca de la crisis del coronavirus que más le interesa conocer?
11. Según el gráfico, indique la opción correcta
12. Indique su distrito
13. Indique qué medios utiliza para informarse
14. Cree que existe algún grupo de edad que esté menos informado
15. Barrio en el que se reside o al que se desplaza habitualmente

Las siguientes variables posiblemente pasarán a ser llamadas en la base de datos, como método para reducir el tamaño del texto, de la siguiente manera:

1. MPFO (Método de Propagación según Fuentes Oficiales)
2. Año
3. Sexo
4. Situación
5. CA (Comunidad Autónoma)
6. Estudios
7. CP (Código postal)
8. Rapidez

9. Transparencia de datos del gobierno central, autonómico y local
10. Información
11. Interpretación de gráficos
12. Distrito
13. Medios para informarse
14. Grupo de edad menos informado
15. Barrio

MPFO (Método de Propagación según Fuentes Oficiales)

-Definición de la variable

MPFO es una variable que recoge datos basados en la opinión de la gente. Esto nos permite saber el punto de vista de las personas respecto a el método de propagación del COVID-19. En la pregunta se especifica que el método debe ser según lo indicado por las fuentes oficiales. Esto nos permite identificar si la gente se ha informado de forma correcta o por el contrario la ha adquirido por bulos o información no oficial.

-Tipo de la variable

Se trata de una variable cualitativa con opciones de respuesta expresadas con texto.

-Sus clases con su código para las variables cualitativas

Las opciones disponibles de respuesta son:

- Propagación a través del aire
- Propagación por gotas al toser y estornudar
- Propagación a través de insectos
- Otra

En la opción 'Otra' damos libertad de respuesta a la persona que está respondiendo el cuestionario.

Se trata de una pregunta que permite marcar más de una opción, puesto que una misma persona puede haber escuchado que el virus se transmite de varias formas y creer que hay varias respuestas posibles.

-Tiempo en el que se medirán

Las respuestas están recogidas entre el 27-28 de abril de 2020. Esta información es importante saberla puesto que el pensamiento de las personas es cambiante en el tiempo, al igual que también lo es la información proporcionada por las fuentes oficiales. Hay que recordar que, en este caso, estamos tratando con un virus que ha aparecido por primera vez y todavía se está investigando, por lo que la información generada puede ser renovada.

-Cualquier otro detalle de interés para la variable

Esta variable ha sido creada con el objetivo de medir como de informada se encuentra la población e incluso poder concluir, en caso de que se hayan informado de una fuente no oficial, el impacto que podrían tener los bulos en las personas.

AÑO (Año de nacimiento)

-Definición de la variable

Año de nacimiento es una variable que permite saber la edad que tiene la persona en cuestión.

-Tipo de la variable

Se trata de una variable cuantitativa discreta puesto que pedimos un número de 4 cifras, que por lo tanto será un entero.

-Su rango en el caso de las variables cuantitativas

Podremos saber el rango de esta variable para la muestra tomada una vez se haya lanzado la encuesta a las personas. Hasta el momento, en fecha 26 de abril podemos decir que el rango, por lógica, debería de pertenecer a un intervalo de personas con edad entre 14 y 90 años, es decir entre 1930 y 2006. En caso de tener respuesta de una persona mayor de 90 años o menor de 14 se aceptaría, simplemente hemos supuesto que lo más normal es que esto no suceda.

-Tiempo en el que se medirán

En este caso, el tiempo en el que se toma la variable no tiene relevancia puesto que la fecha de nacimiento no es un dato cambiante en el tiempo. Si que lo sería la edad de la persona que se puede obtener a partir de esta. No obstante, como la variable es el año de nacimiento, cualquier persona interesada en reutilizar la base de datos podrá obtener la edad actual de la persona haciendo una resta de la fecha en que se encuentre y la fecha en que se tomaron los datos (27-28 de abril de 2020).

-Cualquier otro detalle de interés para la variable

El objetivo por el que se hace esta pregunta es para poder identificar comportamientos según la edad de la persona. Para esto, crearemos una nueva variable 'EDAD' que será cuantitativa discreta también y la recodificaremos de manera que tengamos la información por intervalos de edad.

SEXO

-Definición de la variable

Sexo es una variable que recoge información referente a la condición orgánica de una persona.

-Tipo de la variable

Se trata de una variable cualitativa puesto que las respuestas posibles son de carácter textual.

-Sus clases con su código para las variables cualitativas

Las opciones que se proporcionan como respuesta son:

- Masculino
- Femenino
- Prefiero no decirlo
- Otro

De esta manera se puede responder con total libertad a la pregunta. La opción 'Otro' permite escribir.

-Tiempo en el que se medirán

Esta variable se ha medido entre el 27-28 de abril de 2020.

-Cualquier otro detalle de interés para la variable

El objetivo de esta variable es tener la oportunidad de encontrar tendencias según el sexo de la persona. Por ejemplo, mediante tabulaciones cruzadas.

SITUACIÓN

-Definición de la variable

Situación es una variable que recoge información solamente de los ciudadanos que viven en la Comunidad Valenciana. Esta variable nos permite saber la situación de la persona de manera más concreta.

-Tipo de la variable

Se trata de una variable cualitativa con opciones de respuesta expresadas con texto.

-Sus clases con su código para las variables cualitativas

Las opciones disponibles de respuesta son:

- Vivo en Valencia
- Trabajo en Valencia y me desplazo también ahora de manera habitual
- Me desplazo ahora de manera habitual a Valencia para el cuidado de familiares o enfermos u otra situación
- Vivo en otra población y no voy a Valencia durante la cuarentena

Las tres primeras opciones llevan a una sección donde se pide el código postal y el distrito. La última opción, al no ser de Valencia, pasa directamente a la última sección que puede contestar de manera no obligatoria todas las personas.

-Tiempo en el que se medirán

Las respuestas están recogidas entre el 28-29 de abril de 2020. El tiempo en el que se recogen los datos, en la gran mayoría de los casos, no debería influir en la respuesta de la persona.

-Cualquier otro detalle de interés para la variable

Esta variable ha sido creada con el objetivo de conocer la situación actual en la que se encuentra la persona que vive en la Comunidad Valenciana. Además nos aporta detalles extra y nos permite cruzar información con la encuesta lanzada por el Ayuntamiento de Valencia.

CA (Comunidad Autónoma)

-Definición de la variable

CA es la variable que recoge el lugar de residencia del encuestado a nivel de comunidad autónoma. Esta variable es imprescindible para conocer la percepción de la ciudadanía sobre la rapidez de la gestión de su gobierno autonómico respecto a la crisis del COVID-19.

En la encuesta esta pregunta se presenta como una lista desplegable donde el encuestado puede seleccionar la Comunidad Autónoma correspondiente. La encuesta es a nivel nacional, con lo que contempla regiones que no sean españolas.

Ceuta y Melilla están incluidas en las opciones.

-Tipo de variable

CA es una variable de tipo cualitativa de opción única.

-Clases de la variable cualitativa

CA puede adoptar 19 valores diferentes, cada uno correspondiente con el nombre de las Comunidades Autónomas de España. Las variables carecen de código.

-Tiempo en el que se medirá

Las respuestas a esta pregunta de la encuesta serán recogidas entre los días 27 y 28 de abril de 2020.

-Cualquier otro detalle de interés para la variable

Dependiendo del grado de difusión de la encuesta, existe la posibilidad de que si la frecuencia relativa de un valor (se espera la Comunidad Valenciana) sea muy alto (del 95% o superior) se reduzca el estudio a la percepción de la ciudadanía sobre la rapidez de la gestión respecto a la crisis del COVID-19 en la Comunidad Valencia en vez de España.

ESTUDIOS (Nivel de estudios)

-Definición de la variable

Estudios es la variable que recoge el nivel máximo de estudios del encuestado. Esta variable puede utilizarse para determinar si existe relación entre el nivel de estudios y la capacidad de interpretar información, así como comprender la situación actual y la gestión del Gobierno.

-Tipo de variable

Estudios es una variable de tipo cualitativa de opción única.

-Clases de la variable cualitativa

Estudios puede adoptar los siguientes valores:

- Educación primaria
- Educación Secundaria Obligatoria
- Bachillerato/Ciclos Formativos
- Universitario
- Otro: en este caso el encuestado introduce manualmente su nivel de estudios si no coincide con ninguna de las opciones anteriores

-Tiempo en el que se medirá

Las respuestas a esta pregunta de la encuesta serán recogidas entre los días 27 y 28 de abril de 2020.

-Cualquier otro detalle de interés para la variable

Debido a la brevedad de tiempo para recoger respuestas, se espera que los valores con mayor frecuencia relativa sean “Universitario” y “Bachillerato/Ciclos formativos”

CP (Código postal)

-Definición de la variable

CP es una variable identificadora que recoge el código postal del encuestado si vive o se desplaza para trabajar en Valencia. Esta variable podrá cruzarse posteriormente con otras para realizar mapas y gráficos geográficos.

-Tipo de variable

Estudios es una variable de tipo cualitativa de opción única mediante lista desplegable. Aunque los códigos son numéricos, no representan la medida de ninguna magnitud, con lo que se tratará como cualitativa.

-Clases de la variable cualitativa

Cp puede adoptar los siguientes valores:

- | | | |
|---------|---------|---------|
| • 46001 | • 46011 | • 46021 |
| • 46002 | • 46012 | • 46022 |
| • 46003 | • 46013 | • 46023 |
| • 46004 | • 46014 | • 46024 |
| • 46005 | • 46015 | • 46025 |
| • 46006 | • 46016 | • 46026 |
| • 46007 | • 46017 | • 46035 |
| • 46008 | • 46018 | • 46920 |
| • 46009 | • 46019 | |
| • 46010 | • 46020 | |

-Tiempo en el que se medirá

Las respuestas a esta pregunta de la encuesta serán recogidas entre los días 27 y 28 de abril de 2020.

-Cualquier otro detalle de interés para la variable

En función del valor escogido por el encuestado en esta variable y de la variable Distrito las posibles opciones para la variable Barrio cambiarán.

Rapidez (En una escala del 1 al 5 valore la rapidez con la que el gobierno está tomando medidas contra la pandemia)

-Definición de la variable

La Rapidez es una variable que estudia la opinión de la población acerca de cómo de rápidas y por lo tanto, efectivas, han sido las medidas que ha ido tomando el gobierno central.

Se mide en forma de una escala lineal del 1 al 5 donde el 1 indica que el gobierno ha actuado lentamente en todas sus medidas, mientras que el 5 indica que el gobierno siempre ha ido un paso por delante del virus

-Tipo de variable

La Rapidez es una variable de tipo cualitativa de una única elección.

-Clases de la variable cualitativa

La rapidez puede tomar 5 valores distintos, cada uno con diferente significado:

- **1** El gobierno ha actuado con poca rapidez en todas sus medidas
- **2** El gobierno ha actuado con calma a la hora de aplicar medidas
- **3** El gobierno ha actuado algunas veces de forma rápida
- **4** El gobierno ha actuado bastante bien en general
- **5** El gobierno siempre ha ido un paso por delante del virus con sus medidas

-Tiempo en el que se medirá

Las respuestas a esta pregunta de la encuesta serán recogidas entre los días 27 y 28 de abril de 2020.

-Cualquier otro detalle de interés para la variable

Aunque sea numérica se estudia de forma cualitativa ya que muestra la opinión en una escala del 1 al 5, no obstante a la hora de hacer inferencia se puede tratar como una cuantitativa.

Con esta pregunta pretendemos obtener la opinión de la gente acerca de cómo creen que lo está haciendo el gobierno, centrándonos en la rapidez de sus medidas, algo fácil de valorar

Transparencia de datos del gobierno central, autonómico y local

-Definición de la variable

En este caso utilizamos una variable cuantitativa discreta (5 posibles valores) para que el encuestado valore la transparencia de los datos que el gobierno central, autonómico y local anuncian cada día respecto a la pandemia. Esta variable es de carácter ordinario, siendo 1 el grado de menor transparencia y 5 el máximo, es decir, los 5 valores que puede responder el encuestado se corresponderían cada uno con una valoración:

-Tipo de variable

Variable de tipo cualitativa de una única elección.

-Clases de la variable cualitativa

Puede tomar 5 valores distintos, cada uno con diferente significado:

- **1** El gobierno no es nada transparente
- **2** El gobierno es poco transparente
- **3** El gobierno no es ni poco ni muy transparente
- **4** El gobierno es transparente
- **5** El gobierno es totalmente transparente

Esta variable aparece en el cuestionario en forma de matriz 3x6 por lo que el encuestado puede dar una valoración al gobierno central y otra diferente al autonómico dependiendo la C. Autónoma en la que resida. Por tanto, esta variable realmente se dividirá en términos de análisis en 3, una que corresponderá a la opinión de transparencia sobre el gobierno central, otra que corresponderá a la transparencia de datos en función de la Comunidad Autónoma y otra que corresponderá con el municipio de los encuestados.

Información(¿Cuál es la información acerca de la crisis del coronavirus que más le interesa conocer?)

-Definición de la variable

La Información es una variable que estudia la opinión de las personas sobre los datos que proporciona el gobierno central, cada día alrededor de la misma hora, en las noticias dan datos sobre infectados, fallecidos, ingresados... pero... ¿sobre cuál se preocupan más los ciudadanos?

-Tipo de variable

La Información es una variable cualitativa de una única elección

-Clases de la variable cualitativa

La Información puede tomar estos valores:

- **Fallecidos** interés por nº de fallecidos por coronavirus (diario o en total)
- **Infectados** interés por nº de casos confirmados por coronavirus (diario o en total)
- **Recuperados** interés por nº de pacientes recuperados de coronavirus (diario o total)
- **Hospitalizados (No en UCI)** interés por nº de pacientes ingresados en hospitales (diario o total)
- **Ingresados en UCI** interés por nº de pacientes en estado crítico (diario o total)
- **Otro** Opción para que el encuestado elija su dato de mayor interés si no está en la encuesta

-Tiempo en el que se medirá

Las respuestas a esta pregunta de la encuesta serán recogidas entre los días 27 y 28 de abril de 2020.

-Cualquier otro detalle de interés para la variable

Las variables son independientes entre sí, aunque numéricamente dependan unas de otras, porque se quiere medir el interés de la población por estos datos. Podríamos averiguar en qué información debería centrarse el gobierno a la hora de resaltar datos.

Interpretación de gráficos

Esta variable nos permitirá saber si el encuestado sabe interpretar un gráfico y sacar conclusiones a partir de este. Se trata de una variable cualitativa, en concreto, binomial cuyos valores posibles son 'Sí sabe interpretar gráficos' y 'No sabe interpretar gráficos'. Como en la pregunta de la encuesta hay 4 opciones y solo una es verdadera, el código que utilizaremos es el siguiente:

- A los encuestados que marquen la tercera opción se les asignará el valor 'Sí sabe interpretar gráficos'.

- A los encuestado que escojan cualquiera de las otras 3 opciones se les asignará el valor 'No sabe interpretar gráficos'.

El gráfico que hemos puesto en el cuestionario lo hemos obtenido de una fuente secundaria, en concreto de una noticia publicada por el periódico digital 'Hoy' en su página web el día 23 de abril de 2020.

Distrito (Indique su distrito)

Definición de la variable

El distrito es una variable que estudia el lugar geográfico donde vive la persona que responde a la encuesta, sólo la pueden responder aquellos que vivan o se desplacen a la ciudad de Valencia

Tipo de variable

El distrito es una variable cualitativa de una única elección.

Clases de la variable cualitativa

1. Ciutat Vella	8. Patraix	15. Rascanya
2. l'Eixample	9. Jesús	16. Benicalap
3. Extramurs	10. Quatre Carreres	17. Pobles del Nord
4. Campanar	11. Poblots Marítims	18. Pobles de l'Oest
5. la Saïdia	12. Camins al Grau	19. Pobles del Sud
6. el Pla del Real	13. Algirós	
7. l'Olivereta	14. Benimaclet	

Tiempo en el que se medirá

Las respuestas a esta pregunta de la encuesta serán recogidas entre los días que dure la encuesta.

Cualquier otro detalle de interés para la variable

La variable depende de que el encuestado responda a que vive en la Comunidad Valenciana y además que viva o se desplace a Valencia durante la cuarentena, de la misma forma, depende del distrito que elija el encuestado se le preguntará por un grupo de barrios, esto provoca que la encuesta tenga muchas secciones. El objetivo de preguntar por todo esto es para poder cruzar nuestras respuestas con la base de datos de Valencia.

Medios para informarse

Gracias a esta variable, seremos capaces de conocer el medio más utilizado por los encuestados y sacar conclusiones sobre la validez o calidad de estos medios. La variable es cualitativa nominal, ya que carecen de orden entre sí.

Hay diez opciones que señalan un medio en concreto y una última opción que da posibilidad a respuesta abierta.

La información que obtenga será utilizada tal cual se recoja, no se modificará de ninguna manera para llevar a cabo el análisis.

Grupo de edad menos informado

La variable nos dará a conocer la perspectiva que tiene cada encuestado en cuanto al rango de edad que está menos informado sobre la situación actual del Covid-19. También se contempla la posibilidad de que todos estén igual de informados.

Los rangos de edad son:

- Jóvenes
- Adultos
- Ancianos
- No hay diferencia entre ellos (todos igual de informados)

Esta variable es de tipo cualitativa nominal. En cuanto al tratamiento de esta, seguramente será tipificada para llevar a cabo un estudio cuantitativo sobre esta.

Barrio en el que se reside o al que se desplaza habitualmente

Esta variable nos indicará dónde se concentra nuestra muestra, ya sea por residencia, trabajo u otros. Por tanto, podemos diferenciar según zonas de Valencia.

Por lo tanto, esta variable se trata de una cualitativa nominal, debido a que no hay un rango (orden) determinado entre los barrios.

Datos recogidos: como se almacenan y gestionan

Con el fin de estudiar la percepción de la ciudadanía sobre la transparencia y presentación de datos respecto a la crisis del COVID-19, se van a recoger datos de un formulario de Google (fuente primaria) creado por los investigadores. La encuesta será difundida por redes sociales.

Enlace al cuestionario:

https://docs.google.com/forms/d/e/1FAIpQLSd8v4PzO35ZYIcloVyjPksgCv64A-V_fobk89YpSyDWc-xsNw/viewform

La encuesta es completamente anónima y las respuestas quedan registradas en un documento de hoja de cálculo de Drive. Para su análisis se utilizarán el programa estadístico StatGraphics y el software de gestión de datos Tableau.

Ficha técnica del cuestionario

Tipo de encuesta	Fija, sin periodicidad
Ámbito poblacional	Toda la población
Ámbito geográfico	Territorio Nacional
Periodo de referencia de los resultados	Junio 2020
Periodo de referencia de la información	28 Abril 2020 – 9 Mayo 2020
Tamaño de la muestra	625 casos y 15 variables (tras limpiar la muestra)
Tipo de muestreo	Muestra aleatoria simple. Utilizaremos un conglomerado bietápico estratificado puesto que haremos una división de nuestra muestra según intervalos de edad.
Método de recogida	Encuesta personal y anónima. Vía redes sociales
Error muestral	Error de estimación: 0.04 Nivel de confianza: 95%

Selección de la muestra y tratamiento de los datos

Tras la publicación y difusión de la encuesta por redes sociales y canales oficiales de la Càtedra Govern Obert, nuestra muestra se componía de 625 casos.

Tras un proceso de tratamiento y limpieza de datos eliminamos los casos que cumplían una de las características siguientes:

- Hay un valor anómalo en alguna de las variables (Ejemplo: Año de nacimiento 1200)
- Al menos el 30% de las variables son datos faltantes (Ejemplo: de 10 preguntas contestó 5)
- Hay incongruencias en las respuestas a diferentes preguntas (Ejemplo: nivel de estudios universitario. Año de nacimiento: 2006)

Una vez realizado todo este proceso de limpieza y tratamiento de datos, la muestra que vamos a utilizar en el análisis y la inferencia se compone de 613 casos, en su gran mayoría completos. Es decir, prácticamente todas las respuestas obtenidas al formulario se han catalogado como válidas, ya que no hay indicios suficientes para ser consideradas como anómalas o atípicas

Colectivos con los que se compartirán los resultados

Uno de los propósitos es cruzar datos con el ayuntamiento de Valencia, que de forma paralela ha realizado una encuesta similar. Además, es una encuesta de interés general puesto que trata sobre la pandemia que se está viviendo a nivel mundial, pero en concreto es de interés para la sociedad española. Por este motivo, se compartirá tanto con el ayuntamiento de Valencia como con el resto

de la población. Aportar conocimiento a las personas es uno de los objetivos del trabajo y para ello es esencial dejar los resultados libremente abiertos a cualquiera.

Análisis DAFO

La matriz de análisis DAFO, es una herramienta estratégica de análisis de la situación del proyecto cuyo principal objetivo es ofrecer un claro diagnóstico para poder tomar las decisiones estratégicas oportunas y mejorar en el futuro. Su nombre deriva del acrónimo formado por las iniciales de los términos: debilidades, amenazas, fortalezas y oportunidades. La matriz de análisis DAFO permite identificar tanto las oportunidades como las amenazas que se presentan en nuestro entorno, y las fortalezas y debilidades que muestra nuestro proyecto.

-Oportunidades: representan una ocasión de mejora del proyecto. Las oportunidades son factores positivos y con posibilidad de ser explotados por parte de los integrantes del mismo.

-Amenazas: pueden poner en peligro el trabajo realizado o en menor medida afectar negativamente a los resultados obtenidos. Si identificamos una amenaza con suficiente antelación podremos evitarla o convertirla en oportunidad.

-Fortalezas: Son todas aquellas capacidades y recursos con los que cuentan los participantes del proyecto para explotar oportunidades.

-Debilidades: Son aquellos puntos de los que el proyecto carece o simplemente de aquellos en los que se puede mejorar.

Formulación Síntesis del Análisis

	OPORTUNIDADES	AMENAZAS
Externo	<ul style="list-style-type: none"> - Aportación de datos para el análisis, a partir de la encuesta. - Gran difusión de la encuesta. - Ayuda externa por parte del profesorado. 	<ul style="list-style-type: none"> - Aportación de datos deficientes - Escasa variedad en relación a los datos. - Dificultad para extrapolar a nivel nacional debido a los datos recolectados.
	FORTALEZAS	DEBILIDADES
Interno	<ul style="list-style-type: none"> - Software: Statgraphics, Python, Excel. - Conocimiento para el tratamiento de datos. - Capacidad de crear valor a partir de resultados. 	<ul style="list-style-type: none"> - Ningún colectivo implicado en el proyecto. - Posible falta de conocimientos en algún proceso. - Dificultad de acceso a la información. - Sin presencia offline. - Tiempo acotado.

Explicación:

-Oportunidades: Gracias a la encuesta realizada, podemos disponer de una gran cantidad de datos para realizar el proyecto. Además, esta encuesta ha tenido una muy buena difusión, por lo que ha llegado a mucha gente tanto de la Comunidad Valenciana como de otras autonomías como son Andalucía, Cataluña, etc.

También tenemos la ocasión de solicitar ayuda al profesorado para la realización del estudio, de tal manera que si carecemos de conocimientos en algún campo o nos es necesario ayuda para cómo enfocar el trabajo, podremos disponer de esta figura.

-Amenazas: Lo negativo de depender de factores externos como pueden ser los individuos que realizan la encuesta es que pueden afectar tanto de forma positiva como negativa, es por ello que se pueden recolectar datos que no nos sean de utilidad si su calidad no es positiva. De igual manera, otra posibilidad es que no haya gran variedad en las respuesta a la encuesta, por lo que no tendríamos la oportunidad de realizar ciertos análisis como pueden ser la inferencia y no lleguen a ser tan útiles cómo podrían llegar a ser.

En relación con los problemas anteriores, los componentes del proyecto disponemos mayoritariamente de un círculo social centrado en la Comunidad Valenciana. Esto causa que los datos que recogemos sean de este punto geográfico también, lo que ocasiona dificultad para ejecutar un trabajo en el cual seamos capaces de extrapolar los resultados y conclusiones a toda la población española.

-Fortalezas: Disponemos de Software proporcionado por la Universitat Politècnica de València. En este proyecto utilizamos Statgraphics para el análisis y representación de los datos, Python para la limpieza y modificación de variables y Excel para administrar y recoger los datos que nos proporciona la encuesta. Además, uno de los puntos que aporta validez y verifica que las conclusiones están basadas en un criterio científico viene dado por los conocimientos adquiridos por los integrantes, que pertenecen al grado Ciencia de Datos, en relación con la manipulación de bases de datos así como la interpretación de gráficos o utilización de métodos estadísticos para extrapolar conclusiones.

Por último, cabe añadir que de los datos podemos obtener información pero ésta no es útil si no la convertimos en conocimiento. El conocimiento lo aportan las personas haciendo comparaciones o aplicando la información basándose en un contexto.

-Debilidades: No contamos con ningún colectivo que nos proporcione ayuda y que por tanto pueda añadir valor al proyecto. Las alianzas o colaboraciones aportan nuevos puntos de vista. Además, se

trata de uno de los primeros proyectos que realizamos en profundidad por lo que se ve traducido a una falta de conocimiento en algunos detalles que seguramente se pueden mejorar.

Por otro lado, el proyecto se realiza en período de estado de alarma por lo que cada uno lo realiza desde sus domicilios. La falta de presencia física se transmite en una reducción a nivel de comunicación efectiva.

No podemos optar a todos los datos que nos gustaría tener. Uno de los objetivos del proyecto es poder cruzar información con el ayuntamiento de Valencia, el cual ha lanzado una encuesta similar a la nuestra. No obstante, no tenemos la capacidad para comunicarnos con ellos y poder mejorar el trabajo gracias a los resultados obtenidos por el ayuntamiento, a la misma vez que lo estamos realizando. Sí que podremos comparar resultados una vez estos sean proporcionados de manera pública.

En relación con el tiempo, disponemos de una cantidad limitada. De tal forma que no podemos utilizar grandes cuantías de tiempo para tomar decisiones o efectuar tareas como pueden ser la recolecta de datos, el análisis, etc. Por lo tanto, nos tenemos que acoplar al periodo que tenemos para llevar a cabo el proyecto.

Análisis

El análisis es uno de los apartados que proporciona más valor a nuestro proyecto ya que éste está basado en la realización de una encuesta que requiere de un estudio posterior, tanto nivel individual como cruzando variables.

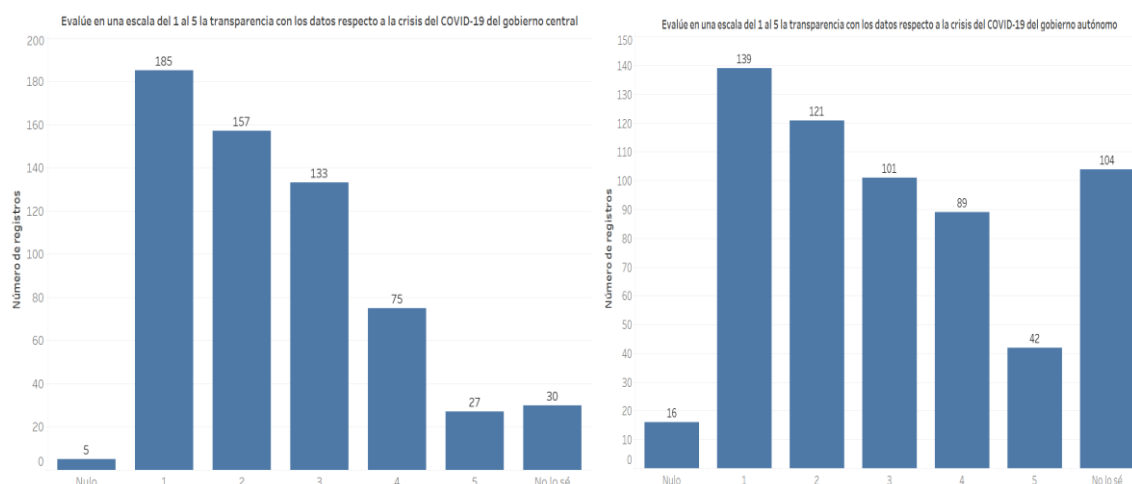
Análisis univariante

En relación con el análisis univariante presentaremos los gráficos que más información nos aporten. Somos consciente de poder representar una misma información de maneras diferentes pero pensamos que no es necesario insertar muchos gráficos con la misma información ya que esto podría derivar en un análisis univariante muy extenso que no aportaría más valor del que mostraremos a continuación.

Transparencia de datos del gobierno central, autonómico y local

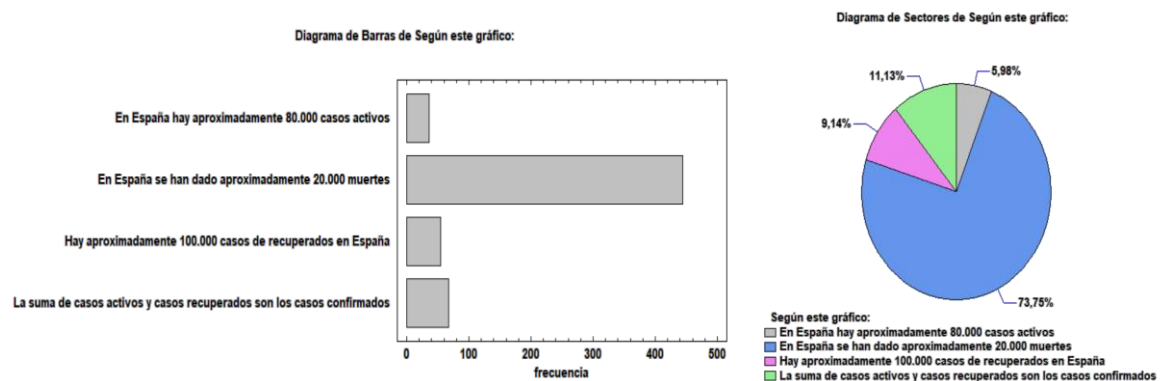
CENTRAL

AUTONÓMICO



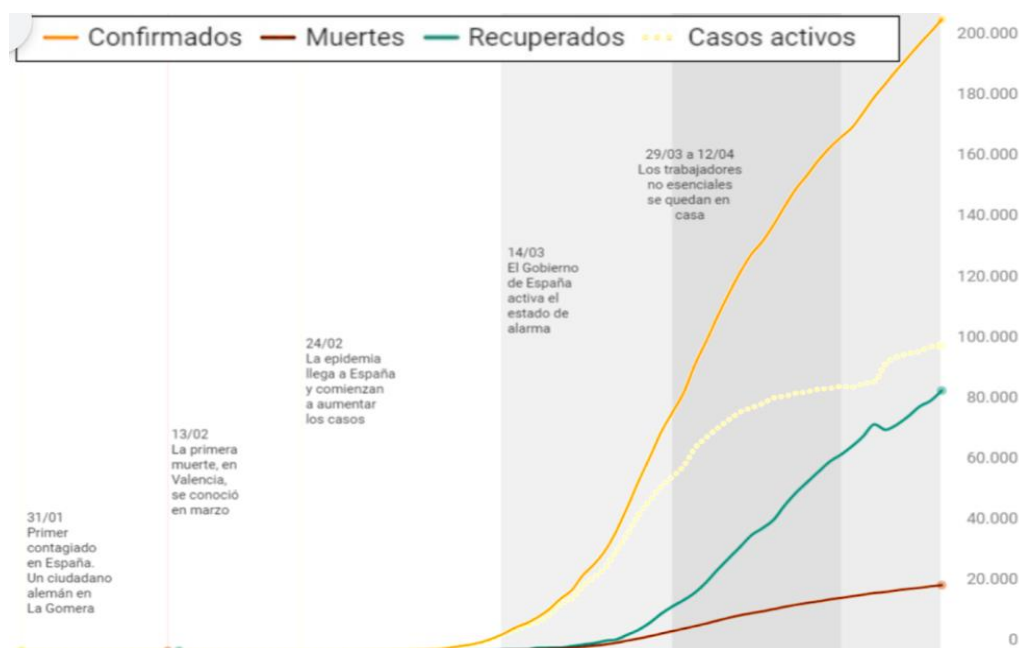
Podemos ver como la opinión de la muestra que representa a la sociedad respecto a la transparencia de los datos tanto del gobierno central como del autonómico tiende a la baja. No obstante, se ve una clara evolución positiva del primero al segundo. Esto se debe a que aumenta el número de personas que no saben el nivel de transparencia del gobierno con los datos de 30 a 104 y a que el número de personas sumando con valoraciones negativas disminuye de 342 (juntando 1 y 2) a 260. Todo lo contrario ocurre con las valoraciones más positivas que pasan de 102 (juntando 4 y 5) a 131.

Interpretación de gráficos

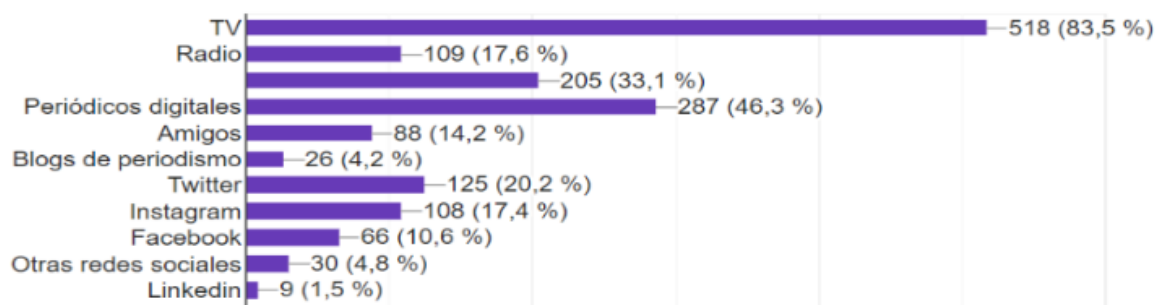


Gracias a estos gráficos podemos observar que la gran mayoría de las personas que han contestado lo han hecho correctamente, concretamente un 73,75%. No obstante, tratándose de un gráfico muy

simple como es el que presentamos en la pregunta del formulario, consideramos que el 26,25%, es decir, más de un $\frac{1}{4}$ de las personas de la muestra es un porcentaje alto que representa que mucha gente no sabe interpretar un sencillo gráfico como el siguiente:



Medios para informarse

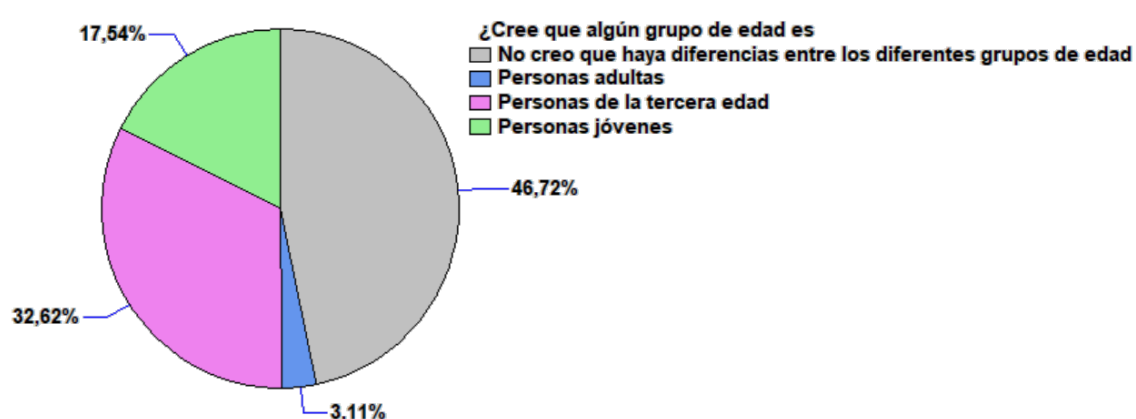


En este gráfico de barras se representa la frecuencia en que la gente ha seleccionado una opción. Hay que remarcar que esta pregunta era 'multiple choice' por lo que la suma total no corresponde

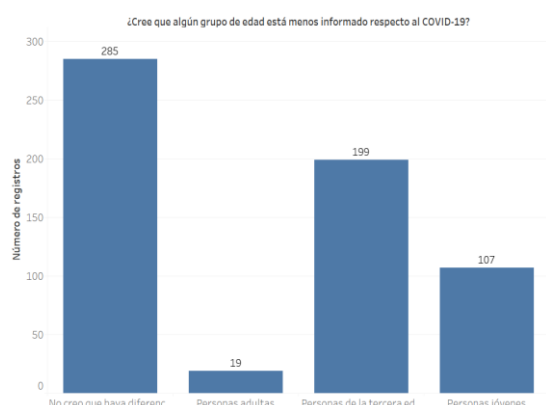
con el total de personas sino con el total de opciones seleccionadas. De esta manera, podemos ver que la frecuencia más elevada corresponde a la televisión con 518 personas, es decir, un 83,5% del total de la muestra y las más bajas a blogs de periodismo. Seguido a la TV se encuentran los periódicos digitales con un 46,3% respecto del total y las fuentes oficiales del gobierno (no aparece el nombre) que corresponde a 205 veces seleccionadas y un 33,1%.

Grupo de edad menos informado

Diagrama de Sectores de ¿Cree que algún grupo de edad es



Según el diagrama de sectores el 46,72% de las personas ha respondido que no considera que haya diferencia entre los diferentes grupos de edad respecto a qué grupo se encuentra menos informado. El siguiente porcentaje es el de las personas de la tercera edad con un 32,62%, siguiéndole un 17,54% que corresponde de los jóvenes. Esta opinión pública se puede deber a suponer que muchas personas de tercera edad no disponen o no utilizan con frecuencia las tecnologías como smartphones que permiten informarse a través de la red.



El siguiente gráfico aporta respecto al anterior una imagen representativa de las frecuencias correspondientes a cada opción marcada. Destacamos que 285 personas no creen que haya ningún grupo de edad menos informado y dentro de los que piensan que si lo hay la tercera edad es el grupo más señalado.

Residentes desplazados

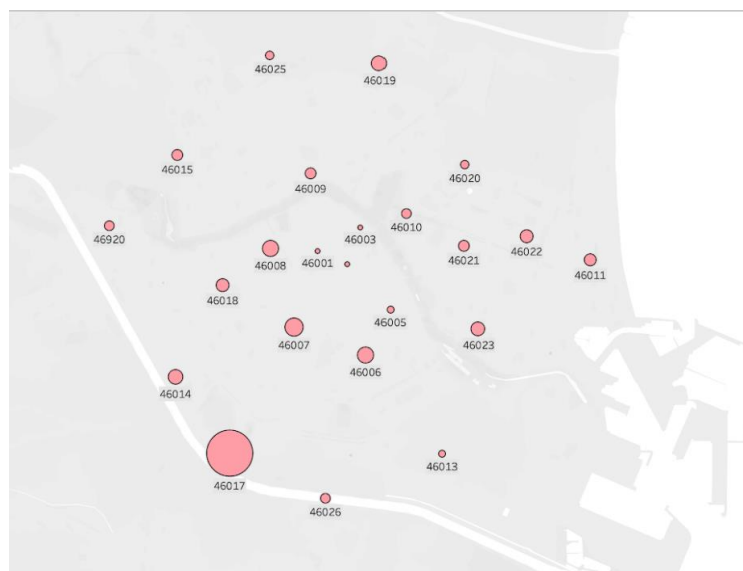


Este gráfico circular trata de los desplazamientos que realiza la población valenciana por la ciudad. Se distinguen 3 tipos de situaciones: No viven ni acuden a Valencia, Viven en Valencia o No viven en Valencia pero acuden habitualmente. Dentro de esta última opción, se puede distinguir entre la persona que acude por trabajo o cuidado de familiares.

Podemos observar que hay dos situaciones principales, las cuales son 'Vivo en Valencia' que es un 43.57% o 'Vivo en otra población y no voy a Valencia durante la cuarentena' que es un 54.77%.

Por lo tanto, podemos afirmar que no hay muchos desplazamientos a Valencia durante la cuarentena.

Códigos postales



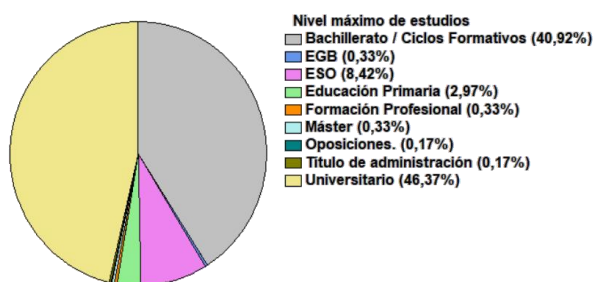
El mapa de códigos postales es muy útil a la hora de distinguir dónde se concentran los individuos de la muestra visualmente.

Por ello, podemos apreciar que la zona con el código postal 46017 contiene a una gran parte de nuestra muestra. Las demás zonas tienen una densidad similar entre ellas.

Se pueden destacar algunas como las zonas 46006, 46007, 46008 que tienen bastante densidad o, por el contrario, zonas como 46001, 46005 o 46003 que tienen una concentración de individuos ínfima.

Nivel máximo de estudios

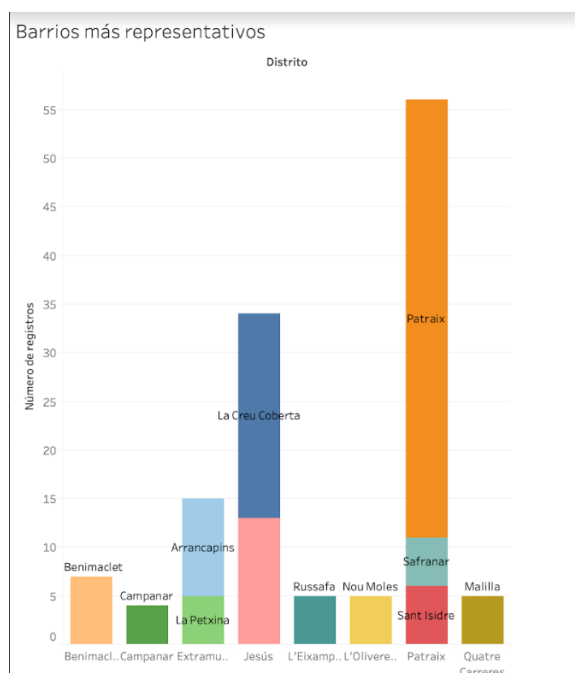
Diagrama de Sectores de Nivel máximo de estudios



Según este diagrama de sectores, en el que tratamos de identificar el nivel máximo de los encuestados, podemos afirmar que hay dos niveles que predominan sobre los demás:
'Universitario' con 46.37%
'Bachillerato/Ciclo Formativos' con 40.92%.

Por ello, podemos saber que casi la mitad de los encuestados tienen estudios superiores, mientras que la otra mitad no.

Distritos-Barrios de los encuestados

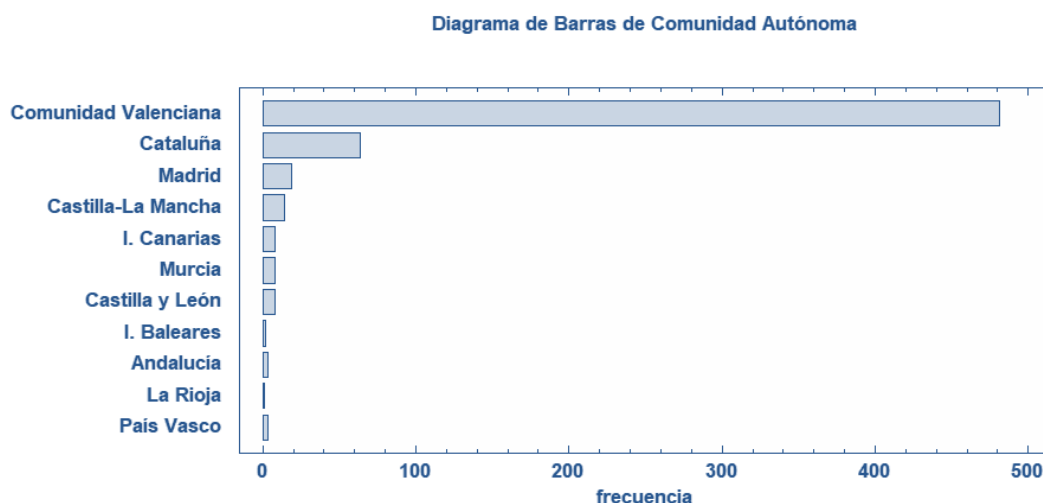


Con este diagrama de barras, conseguimos apreciar en qué distritos y, más en concreto, en qué barrios se concentran los encuestados de manera visual.

Los casos más llamativos en cuanto a barrios son los de 'Patraix', 'La Creu Coberta' y como tercer caso tenemos el barrio de 'Arrancapins' aunque no es tan significativos como los dos anteriores.

En cuanto a los distritos, podemos señalar que los de mayor densidad son 'Patraix' y 'Jesús', en los cuales se encuentran dos de los barrios anteriormente nombrados. También podríamos indicar 'Extramurs' como otro distrito importante.

Distritos-Barrios de los encuestados

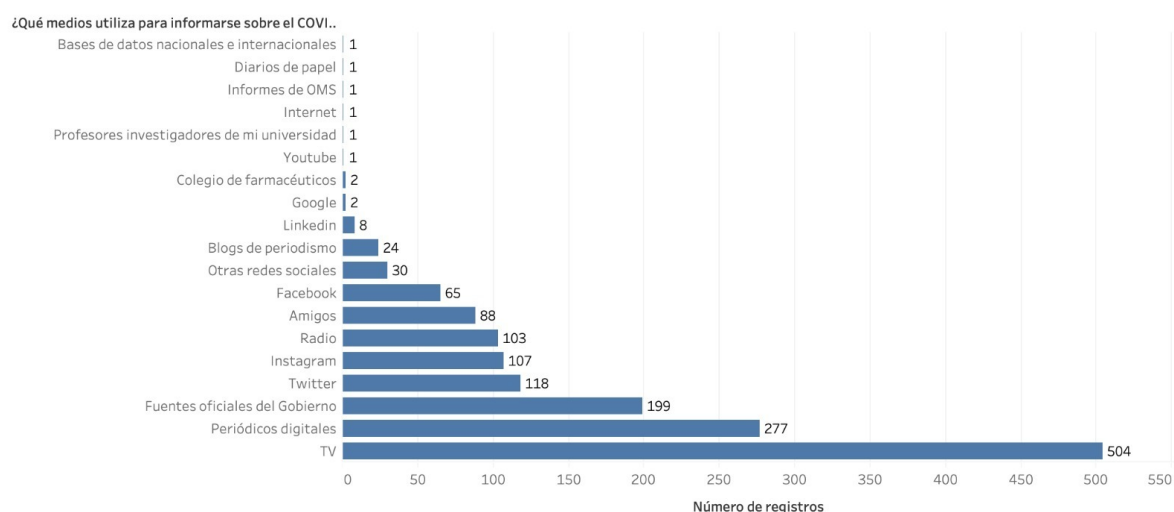


Debido a la gran difusión que ha tenido la encuesta por España, podemos saber en qué CCAA hay encuestados y qué cantidad de estos hay. Por ello, usamos este diagrama de barras en el que se aprecia la Comunidad Valenciana como principal, después tenemos a Cataluña y en tercer puesto tenemos a Madrid.

Tenemos individuos de 11 CCAA en total, lo cual puede ser de gran ayuda a la hora de hacer inferencia a nivel estatal.

Medios de Información más utilizados

Medios Información Más Utilizados

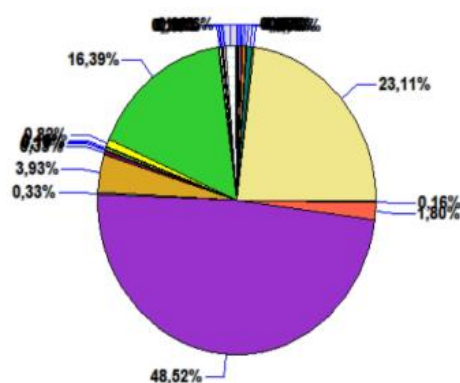


Por último, hacemos uso de un diagrama de barras para analizar qué medios son los más utilizados para informarse uno y para saber el tipo de fuente.

Observamos que hay una gran variedad de medios pero los más usuales son 'TV', 'Periódicos Digitales', 'Fuentes oficiales de Gobierno' y el conjunto de redes sociales. Es por ello que podemos afirmar el alto uso de dispositivos electrónicos para informarse, las clásicas fuentes de información pasan a segundo plano, como son el periódico y la radio.

Un punto a tener en cuenta en este análisis es que los individuos podían señalar más de una opción, es por ello que el total de todos ellos es mayor que la cantidad de gente encuestada.

¿Qué información es más interesante para la sociedad?

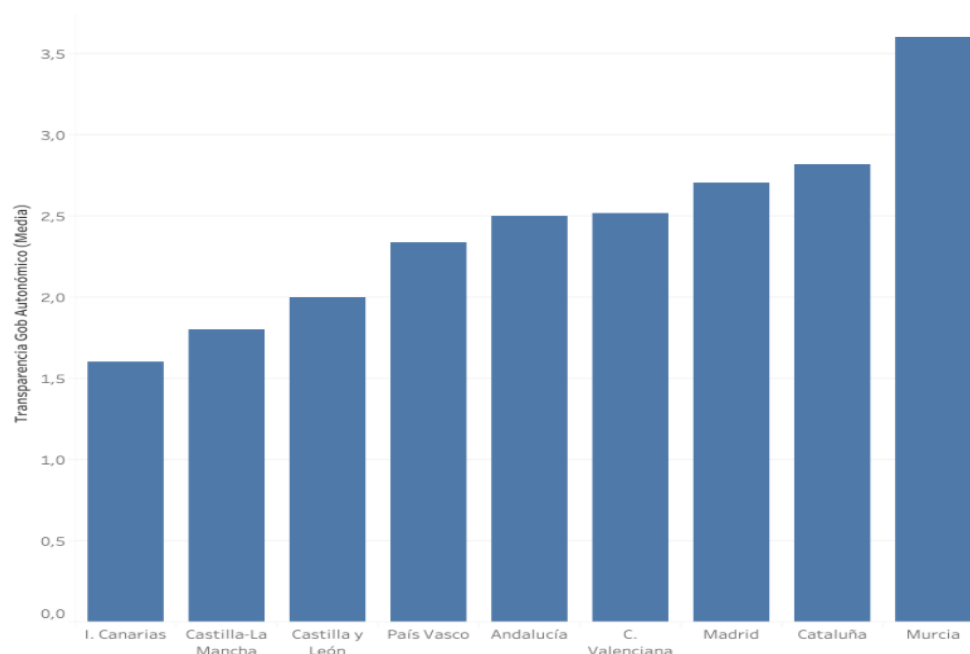


El **48,52%** corresponde a 'Infectados', el **23'11%** a 'Fallecidos' y el **16,39%** a 'Recuperados'.

Observamos que el dato de más interés es el relacionado con los infectados. Suponemos que este resultado es debido a que esta cifra es la más significativa para marcar la evolución de la pandemia.

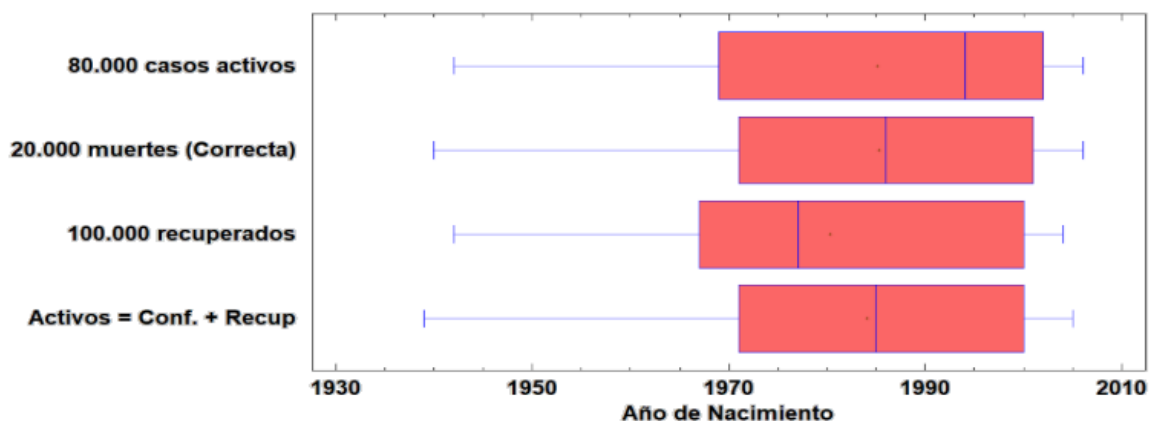
Análisis bivalente

-Transparencia de datos según la Comunidad Autónoma



En el siguiente gráfico podemos ver la media de las respuestas dadas en la pregunta transparencia de datos del gobierno autonómico separada por Comunidades Autónomas, de esta manera podemos saber la opinión de las personas respecto a su propia comunidad. Debemos aclarar que tenemos más datos la Comunidad Valenciana (481 respuestas) y de Cataluña (64 respuestas) por lo que son las más representativas. Por tanto, son las que nos proporcionan más valor. En el caso de la Comunidad Valenciana vemos que la opinión se encuentra dividida y la media es 2'5, es decir, levemente menor al 3 que significa que el gobierno no es ni muy poco transparente ni demasiado, sino que su transparencia es normal, pero en este caso es normal con tendencia negativa. Por parte de Cataluña el resultado es ligeramente superior a este 2'5, es decir, 2'8, un resultado también normal que representa que los catalanes que respondieron el formulario tiene una opinión ni negativa ni positiva sobre la transparencia de datos en su comunidad.

Interpretación de edad según la edad

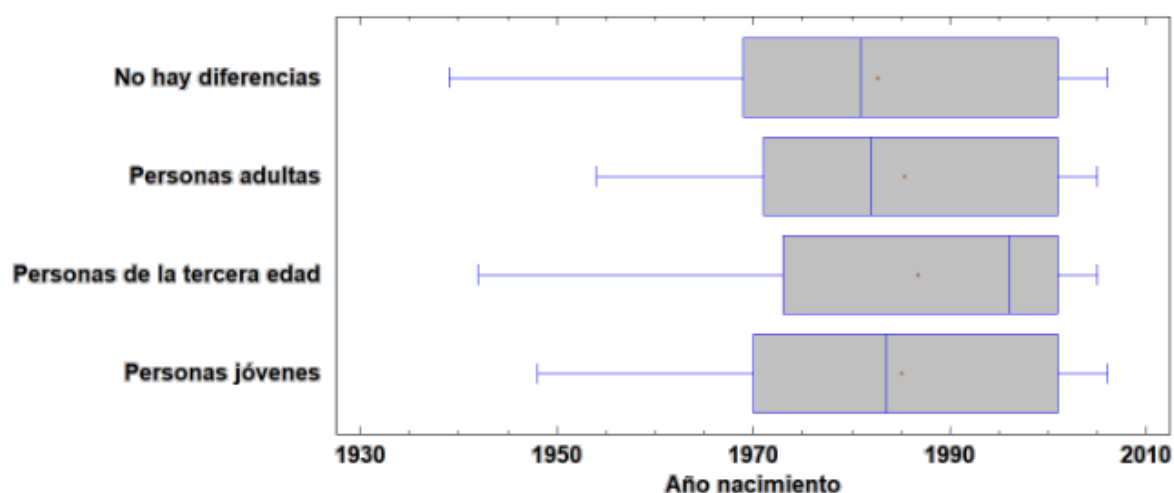


En el siguiente gráfico de caja y bigotes múltiple vemos las respuestas a la pregunta de interpretación del gráfico según el año de nacimiento, esto se puede traducir a la interpretación según la edad. A nivel de rango en casi todos los casos es similar y se encuentra entre [1940 - 2000]. Nos centraremos en la media y donde encontramos el 50% de los datos centrales. Este último para la opción correcta se encuentra entre [1972 - 2002], es decir, el 50% central dentro de las personas que han respondido bien está entre 18 y 48 años. Para dos de las opciones incorrectas este porcentaje se encuentra entre [1966-2004], [1966-2000] y [1972-2000] para las opciones '80.000 casos activos', '100.000 recuperados' y 'activos = confirmados + recuperados', respectivamente. Esto indica que de manera leve hay más personas que contestan de manera incorrecta que tienen más edad. Lo podemos ver comparando medias y medianas con la siguiente tabla.

Grafico	Recuento	Promedio	Mediana
En España hay aproximadamente 80.000 casos activos	36	1985,11	1994,0
En España se han dado aproximadamente 20.000 muertes	441	1985,26	1986,0
Hay aproximadamente 100.000 casos de recuperados en España	55	1980,27	1977,0
La suma de casos activos y casos recuperados son los casos confirmados	66	1984,11	1985,0
Total	598	1984,66	1985,0

Según indica el promedio de la opción correcta es de 1985, es decir, la media de las personas que contestado correctamente es de 35 años. Para las opciones '80.000 casos activos', '100.000 recuperados' y 'activos = confirmados + recuperados', es de 1985, 1980 y 1984, es decir, si hacemos la media de estas 3, sería 1983 o su equivalencia en edad, 37 años. Esto muestra la ligera tendencia a contestar de manera equivocada según aumenta la edad.

Grupo menos informado según la edad



Como podemos ver, el siguiente gráfico sí que tiene diferencias más significativas a nivel de rango. El rango con mayor diferencia es el correspondiente a las personas que no marcan ningún grupo de edad y piensan que no existe relación entre la edad y estar informado erróneamente, esto se debe también a que es la opción más seleccionada. Además, aquellos que tienen entre 78 y 83 años sólo han marcado esta opción y no se han decantado por ningún grupo de edad. Dentro de aquellos que señalan a los adultos como los menos informados encontramos un rango entre [1954-2001] o entre 19 y 66 años. Aquellos que creen que la tercera edad está menos informada nacieron entre [1942-2000] o entre 20 y 78 años. Este dato es significativo porque incluso personas que tienden a la tercera edad se han incluido como grupo de edad menos informada. Respecto a los jóvenes, el rango se muestra entre [1947-2002] o entre 18 y 73 años.

<i>Grupo Edad Desinformado</i>	<i>Recuento</i>	<i>Promedio</i>	<i>Mediana</i>
No creo que haya diferencias entre los diferentes grupos de edad	285	1982,62	1981,0
Personas adultas	19	1985,37	1982,0
Personas de la tercera edad	196	1986,7	1996,0
Personas jóvenes	106	1985,14	1983,5
Total	606	1984,47	1985,0

La medias no muestran una diferencia significativa pero sí que muestran que ligeramente que el promedio de personas más jóvenes de todas las opciones piensa que los menos informados son las personas de tercera edad. En cambio el promedio de personas con más edad de todas las posibles opciones no piensa que sean los jóvenes sino que no se decantan por ningún grupo y consideran que no existe relación. En modo resumen, los más jóvenes señalan a las personas de tercera edad como menos informados y las personas con más edad no señalan a nadie.

Inferencia sobre la población

Métodos de propagación

Esta pregunta del cuestionario era 'multiple choice' por lo que la persona podía elegir más de una opción. No obstante, sólo existe una opción correcta, es decir, hasta el momento sólo se ha podido afirmar por fuentes oficiales como la Organización Mundial de la Salud que el virus se propaga por gotas al toser y estornudar. El resto de las opciones son incorrectas. En caso de que la persona responda más de una opción se considerará que la persona ha respondido erróneamente, puesto que aunque dentro de esas opciones se incluya la correcta se está seleccionando alguna opción que todavía no ha sido validada por las fuentes de información fiables.

Según el análisis univariante de esta pregunta, el 96,4% de las personas han seleccionado la opción de propagación por gotas y estornudar, lo que demuestra que prácticamente toda la muestra es consciente de cómo se transmite el COVID-19. No obstante, la pregunta fue creada con la intención de averiguar el porcentaje de población que había confiado en seleccionar alguna opción no validada por fuente principal de información oficial al respecto de esta pandemia, es decir, la OMS.

La muestra refleja las siguientes proporciones, el 65,03% ha marcado únicamente la opción correcta, el 34,97% restante ha marcado otras opciones. A partir de estos datos haremos inferencia con el objetivo de encontrar información no sólo de la muestra, sino de la población.

A continuación, realizaremos las operaciones que contribuyen a encontrar un intervalo de confianza del 95% para la proporción de personas que han seleccionado alguna opción diferente a la correcta. La fórmula utilizada:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- \hat{p} es la proporción que nos ofrece la muestra
- n es el tamaño de la muestra
- z es el estadístico normal estandarizado en valor absoluto para el α dado dividido entre dos. En este caso $\alpha = 0,05$. Indica que, para la distribución especificada, -1,95997 es el valor más grande tal que la probabilidad de no exceder -1,95997 es menor ó igual a 0,025. Luego $z = 1,96$ en valor absoluto.

Luego, el intervalo será:

$$[0,3497-1,96* \sqrt{((0,3497*0,6503)/612)}, 0,3497+1,96* \sqrt{((0,3497*0,6503)/612)}]$$

Simplificando:

$$[0,311905 , 0,388953]$$

Con un intervalo de confianza del 95% podemos saber que la proporción de personas que han respondido de forma incorrecta esta pregunta oscila entre el 31,1905% y 38,8953% en la población según los datos de la muestra.

Para hacernos una idea veremos en la sociedad española entre qué valores se encuentra este intervalo. Si suponemos que la población española es de 47.000.000 de personas, el intervalo quedaría de la siguiente manera:

$$[14.659.535 , 18.280.791]$$

Entre 14,65 y 18,28 millones de personas en la sociedad española habría interpretado algún ‘bulo’ en referencia a los métodos de propagación como información válida.

A continuación, intentaremos demostrar que existe diferencia entre la probabilidad de que una persona conteste bien o no. Para eso partiremos de una hipótesis nula igual a 0,5, lo que significa que la probabilidad que se responda mal es la misma de que se responda bien.

$$H_0 = 0,5$$

$$H_1 \neq 0,5$$

Utilizaremos p-valor para evidenciar esto. Para ello el primer paso es calcular el estadístico de contraste z para nuestra hipótesis nula.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Siendo $p_0 = H_0 = 0,5$

Luego,

$$z = (0,3497 - 0,5)/ \sqrt{((0,5*0,5)/612)} = -7,4364$$

La probabilidad de encontrar un número menor a -7,4364 en una distribución normal estandarizada es de 5,2125E-14.

$$p\text{-Valor} = 5,2125E-14 * 2 = 1.0425E-13$$

Puesto que $p\text{-Valor} < \alpha$, rechazamos la hipótesis nula, es decir, existe diferencia entre la probabilidad de que una persona conteste bien o no.

Los valores en los que se aceptará la hipótesis nula para un $\alpha = 0,5$, serán aquellos que se encuentren en el intervalo calculado. Un ejemplo podría ser:

$$H_0 = 0,35$$

$$H_1 \neq 0,35$$

Pasamos a calcular z :

$$z = (0,3497 - 0,35) / \sqrt{(0,35 \cdot 0,65) / 612} = -0,01$$

La probabilidad de encontrar un número menor a $-7,4364$ en una distribución normal estandarizada es de $0,493814$.

$$p\text{-Valor} = 0,493814 \cdot 2 = 0,987628$$

Puesto que $p\text{-valor} > \alpha$, no hay evidencias para rechazar la hipótesis nula para $\alpha = 0,05$.

Fuente: http://www.econ.upf.edu/~satorra/dades/moore_08.pdf

Análisis de la interpretación del gráfico de la población

Esta pregunta del cuestionario era tipo test, es decir los encuestados sólo podían escoger una opción de las 4 permitidas. De la misma forma, sólo había una opción correcta, ya que había 3 respuestas no relacionadas con el gráfico, el objetivo era valorar si los encuestados sabían interpretar un gráfico de frecuencias básico relacionado con el COVID-19 en España, un gráfico que perfectamente podría salir en la televisión o en un periódico.

La muestra refleja las siguientes proporciones, el 73.75% ha marcado la opción correcta, el 26.25% restante ha marcado opciones incorrectas. A partir de estos datos haremos inferencia con el objetivo de encontrar información no sólo de la muestra, sino de la población.

A continuación, realizaremos las operaciones que contribuyen a encontrar un intervalo de confianza del 95% para la proporción de personas que han seleccionado alguna opción diferente a la correcta. La fórmula utilizada:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- \hat{p} es la proporción de acierto que nos ofrece la muestra, indica (personas que han acertado la pregunta) / (personas que han contestado la pregunta)

- n es el tamaño de la muestra
- z es el estadístico normal estandarizado en valor absoluto para el α dado dividido entre dos. En este caso $\alpha = 0,05$. Indica que, para la distribución especificada, -1,95997 es el valor más grande tal que la probabilidad de no exceder -1,95997 es menor ó igual a 0,025. Luego $z = 1,96$ en valor absoluto

Luego, el intervalo será:

$$[0,7375 - 1,96 * \sqrt{((0,7375 * 0,2625)/620)}, 0,7375 + 1,96 * \sqrt{((0,7375 * 0,2625)/620)}]$$

Simplificando:

$$[0,6822, 0,7531]$$

Con un intervalo de confianza del 95% podemos saber que la proporción de personas que han respondido de forma correcta esta pregunta oscila entre el 68,22% y 75,31% en la población según los datos de la muestra.

Para hacernos una idea veremos en la sociedad española entre qué valores se encuentra este intervalo. Si suponemos que la población española es de 47.000.000 de personas, el intervalo quedaría de la siguiente manera:

$$[32.063.400, 35.395.700]$$

Entre 32.06 y 35.39 millones de personas en la sociedad española habría interpretado correctamente el gráfico, si nuestra muestra fuera representativa.

A continuación, intentaremos demostrar que existe diferencia entre la probabilidad de que una persona conteste bien o no. Para eso partiremos de una hipótesis nula igual a 0,5, lo que significa que la probabilidad que se responda mal es la misma de que se responda bien.

$$H_0 = 0,5$$

$$H_1 \neq 0,5$$

Utilizaremos p-valor para evidenciar esto. Para ello el primer paso es calcular el estadístico de contraste z para nuestra hipótesis nula.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Siendo $p_0 = H_0 = 0,5$

Luego,

$$z = (0,7375 - 0,5) / \sqrt{((0,5 * 0,5) / 620)} = 10.84$$

La probabilidad de encontrar un número mayor a 10.84 en una distribución normal estandarizada es de 1,87437E-26.

$$p\text{-Valor} = 1,87437E-26 * 2 = 3.74874E-26$$

Puesto que $p\text{-Valor} < \alpha$ (, rechazamos la hipótesis nula, es decir, existe diferencia entre la probabilidad de que una persona conteste bien o no.

Los valores en los que se aceptará la hipótesis nula para un $\alpha = 0,5$, serán aquellos que se encuentren en el intervalo calculado. Un ejemplo podría ser:

$$H_0 = 0,7$$

$$H_1 \neq 0,7$$

Pasamos a calcular z:

$$z = (0,7375 - 0,7) / \sqrt{((0,7 * 0,3) / 620)} = 0.961$$

La probabilidad de encontrar un número menor a 0.961 en una distribución normal estandarizada es de 0.16.

$$p\text{-Valor} = 0.961 * 2 = 1.9234$$

Puesto que $p\text{-valor} > \alpha$, no hay evidencias para rechazar la hipótesis nula para $\alpha = 0,05$.

Con statgraphics podemos determinar si esta variable es independiente del sexo o de la edad de los encuestados, es decir, si guarda o no alguna relación su respuesta con su sexo o edad con la prueba chi-cuadrado.

RELACIÓN CON EL SEXO:

Puesto que el valor-P (0,2557) es mayor o igual que 0,05, no se puede rechazar la hipótesis de que filas y columnas son independientes con un nivel de confianza del 95,0%.

Por lo tanto, el valor observado de “Según este gráfico:” , para un caso en particular, pudiera no tener relación con su valor en “ Indique su sexo”.

Conclusión: Puede no tener relación.

Nota: tener en cuenta que hay un sexo que es “no quiero decirlo” con muy pocos casos que puede alterar el resultado

RELACIÓN CON LA EDAD

Puesto que el valor-P (0,0013) es menor que 0,05, se puede rechazar la hipótesis de que filas y columnas son independientes con un nivel de confianza del 95,0%.

Por lo tanto, el valor observado de “Según este gráfico:” , para un caso en particular, está relacionado con su valor en “Indique su año de nacimiento”

Conclusión: Puede haber relación

Nota: Tener en cuenta que la edad / fecha de nacimiento es una variable con muchos valores diferentes, y por lo tanto algunos con muy pocas muestras por lo que puede alterar el resultado.

Análisis de grupos de edad menos informados

Esta pregunta del cuestionario era tipo test, es decir los encuestados sólo podían escoger una opción de las 4 permitidas. En este caso cualquier respuesta era válida porque estamos valorando la opinión de las personas sobre qué grupos de edad piensan que están menos informados sobre el coronavirus, tenemos los 3 grandes grupos (jóvenes, adultos y de 3a edad) y una opción por si piensan que no hay diferencia.

La muestra refleja las siguientes proporciones, el 46.72% piensa que no hay diferencia, el 3.11% piensa que son las personas adultas, el 32.62% piensa que son las personas de 3a edad y por último el 17.54% piensa que son las personas jóvenes.

A partir de estos datos haremos inferencia con el objetivo de encontrar información no sólo de la muestra, sino de la población.

A continuación, realizaremos las operaciones que contribuyen a encontrar un intervalo de confianza del 95% para cada una de las opciones posibles, la fórmula utilizada:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- \hat{p} es la proporción de acierto que nos ofrece la muestra, indica (personas que han respondido a esa opción) / (personas que han contestado la pregunta)
- n es el tamaño de la muestra

- z es el estadístico normal estandarizado en valor absoluto para el α dado dividido entre dos. En este caso $\alpha = 0,05$. Indica que, para la distribución especificada, -1,95997 es el valor más grande tal que la probabilidad de no exceder -1,95997 es menor ó igual a 0,025. Luego $z = 1,96$ en valor absoluto

Luego, el intervalo para cada opción será:

“No creo que haya diferencias entre grupos de edad”

$$\hat{p} = 0.4672$$

$$n = 610$$

$$z = 1.96$$

$$[0.4672 - 1.96 * \sqrt{((0.4672 * 0.5328) / 610)}, 0.4672 + 1.96 * \sqrt{((0.4672 * 0.5328) / 610)}]$$

$$[0,436606, 0,515793]$$

Población de España:

$$[20.520.200, 24.242.271]$$

“Personas adultas”

$$\hat{p} = 0.0311$$

$$n = 610$$

$$z = 1.96$$

$$[0.0311 - 1.96 * \sqrt{((0.0311 * 0.9689) / 610)}, 0.0311 + 1.96 * \sqrt{((0.0311 * 0.9689) / 610)}]$$

$$[0.017324, 0.044875]$$

Población de España:

$$[814.228, 2.109.125]$$

“Personas de la tercera edad”

$$\hat{p} = 0.3262$$

$$n = 610$$

$$z = 1.96$$

$$[0.3262 - 1.96 * \sqrt{((0.3262 * 0.6738) / 610)}, 0.3262 + 1.96 * \sqrt{((0.3262 * 0.6738) / 610)}]$$

$$[0.288995, 0.363404]$$

Población de España:

[13.582.765 , 17.080.024]

“Jóvenes”

$$^p = 0.1754$$

$$n = 610$$

$$z = 1.96$$

$$[0.1754 - 1.96 * \sqrt{(0.1754 * 0.8246) / 610}, 0.1754 + 1.96 * \sqrt{(0.1754 * 0.8246) / 610}]$$

$$[0.145219 , 0.205580]$$

Población de España:

$$[6.825.312 , 9.662.287]$$

Con Statgraphics podemos determinar si esta variable es independiente del sexo o de la edad de los encuestados, es decir, si guarda o no alguna relación su respuesta con su sexo o edad. Con la prueba chi-cuadrado

RELACIÓN CON EL SEXO:

Puesto que el valor-P (0.1802) es mayor o igual que 0,05, no se puede rechazar la hipótesis de que filas y columnas son independientes con un nivel de confianza del 95,0%.

Por lo tanto, el valor observado de “¿Cree que algún grupo de edad está menos informado?” , para un caso en particular, pudiera no tener relación con su valor en “ Indique su sexo”.

Conclusión: Puede no tener relación.

Nota: tener en cuenta que hay un sexo que es “no quiero decirlo” con muy pocos casos que puede alterar el resultado.

RELACIÓN CON LA EDAD:

Puesto que el valor-P (0.3483) es mayor o igual que 0,05, no se puede rechazar la hipótesis de que filas y columnas son independientes con un nivel de confianza del 95,0%.

Por lo tanto, el valor observado de “¿Cree que algún grupo de edad está menos informado?” , para un caso en particular, pudiera no tener relación con su valor en “Indique su año de nacimiento”.

Conclusión: Puede no tener relación.

Nota: Tener en cuenta que la edad / fecha de nacimiento es una variable con muchos valores diferentes, y por lo tanto algunos con muy pocas muestras por lo que puede alterar el resultado.

Conclusiones objetivas de la transparencia de datos sobre el COVID-19

Uno de los objetivos del trabajo es comparar la opinión de la población con aquello que ocurre en la realidad. Por este motivo, hemos buscado que datos nos proporciona el gobierno central y las Comunidades Autónomas ya que éstas son las responsables de transmitir los datos a la unidad central del gobierno. Por consiguiente, presentaremos las conclusiones sobre esta búsqueda. El proceso ha incluido la descarga de todas las bases de datos que proporcionaba cada web oficial visitada para poder valorar las variables que contenía. Por tanto, no sólo valoraremos la transparencia sino la calidad de las bases de datos. Además, explicaremos detalladamente los datos que ofrece el gobierno central, así como los gráficos que proporciona. Este será el único caso en el que detallaremos, es decir, para las Comunidades Autónomas pasaremos directamente a explicar las conclusiones. Esto se hace para evitar que el documento sea demasiado extenso. No obstante, se pueden buscar las variables y las bases de datos en los enlaces a las webs que se encuentran en el texto.

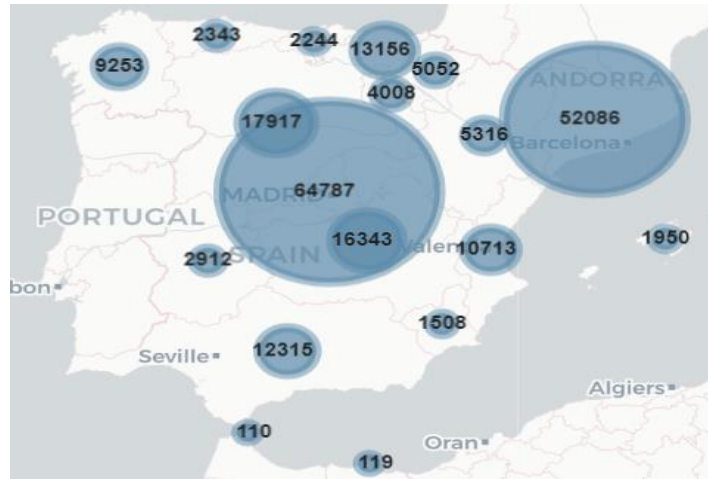
El gobierno ofrece la siguiente web, [COVID-19](#), con el objetivo de mostrar la evolución de la pandemia. Esta página ha sido mejorada con el tiempo. En los inicios de esta situación, se mostraba una tabla con el total de casos acumulados en la comunidad autónoma y los casos cada 24h. Además, ofrecían un enlace para descargar los datos actualizados cada día. En esta base de datos se ofrecía las siguientes variables:

- CCAA
- Fecha
- Casos
- PCR+
- TestAc+
- Hospitalizados
- UCI
- Fallecidos
- Recuperados

Cabe especificar las diferencias entre las variables 'Casos', 'PCR+', 'TestAc+'. En un principio, se registraban todos los casos nuevos en una variable (Casos). Después del 15 de abril, se dividió esta variable en dos variables nuevas, es decir, la variable casos es la suma de las variables PCR+ y TestAc+. El primero nos muestra si la persona tiene actualmente el Covid-19 y el segundo representa si el individuo ha generado anticuerpos de manera que se puede suponer que el Covid-19 ha estado en su cuerpo. Antes del 15 de Abril se consideraban confirmados solo los positivos por PCR, aparentemente no es el caso, por lo que se sustituyeron estos valores por la suma de PCR+ y TestAc+ (Tests rápidos de anticuerpos). Es decir, la suma de ambos nos dice el total de positivos.

Tras la actualización, la página se ha renovado positivamente ofreciendo más información y más gráficos. De esta manera, permite que de manera visual se pueda saber la evolución del COVID-19. A continuación, especificaremos qué información nos ofrecen.

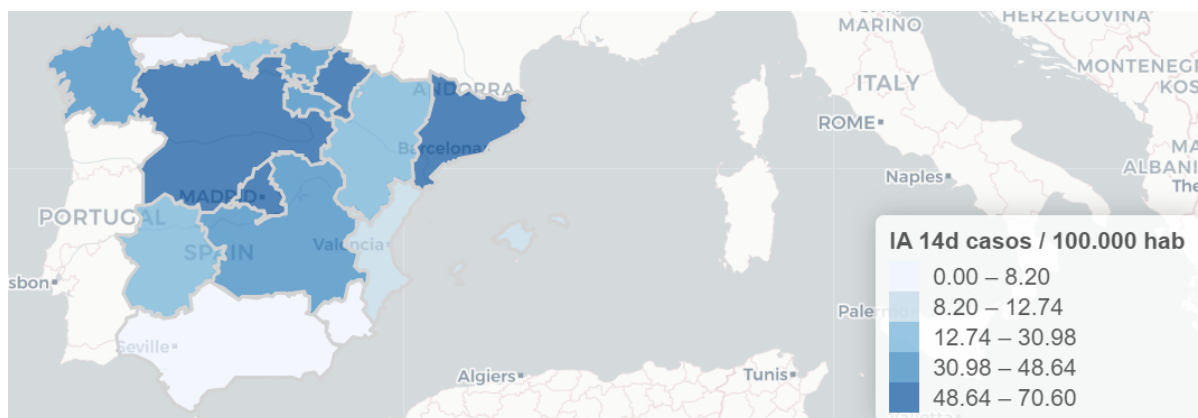
-Casos acumulados por Comunidad Autónoma



-Casos últimas 24h



- Incidencia acumulada últimos 14 días por 100 mil habitantes

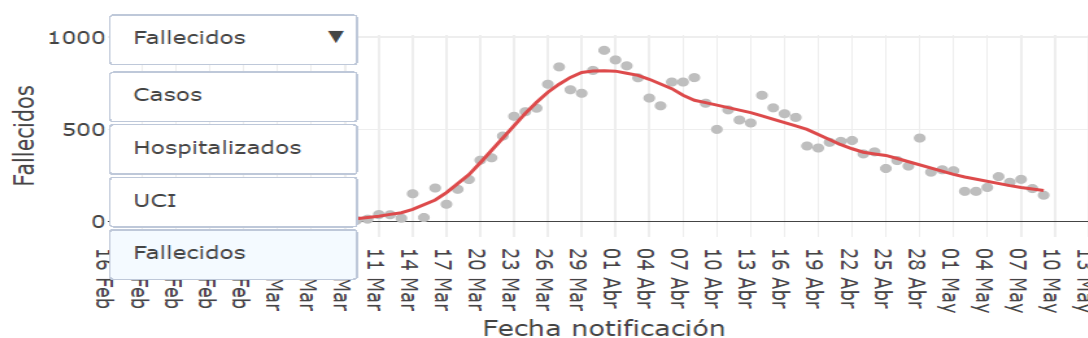


Cifras:

- Casos totales
- Hospitalizados totales
- Recuperados totales
- UCI totales
- Fallecidos totales
- Casos últimas 24 horas
- Hospitalizados últimas 24 horas
- Recuperados últimas 24 horas
- UCI últimas 24 horas
- Fallecidos últimas 24 horas

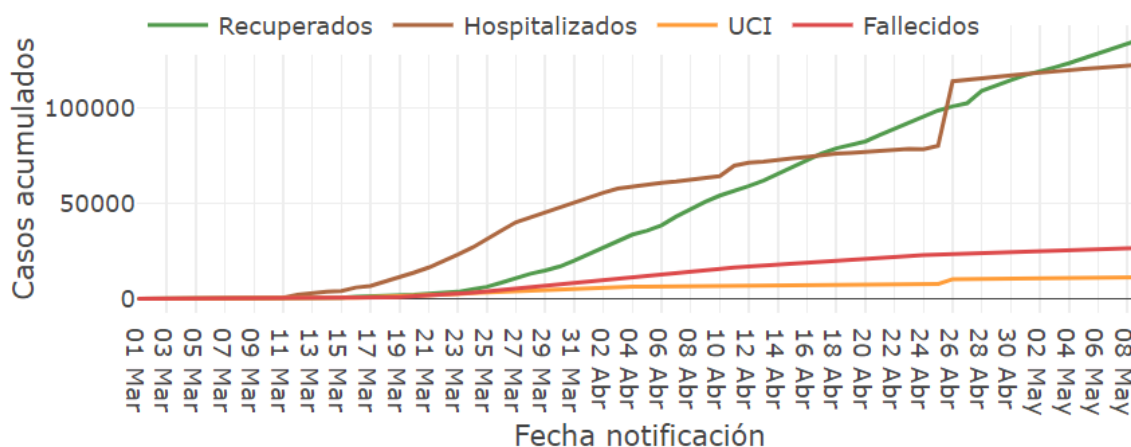
Estos datos se actualizan cada día sobre las 11:00 - 12:30.

Evolución diaria según nivel de gravedad (datos agregados Min. Sanidad)



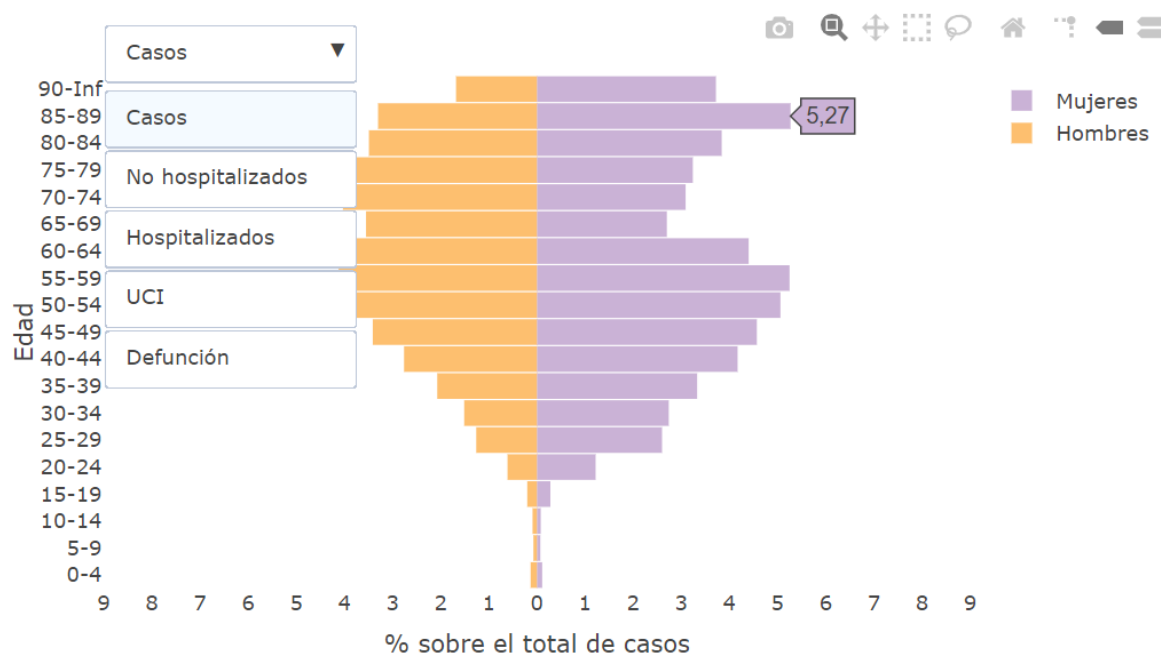
Permite ver la evolución de los casos, hospitalizados, fallecidos y personas en UCI por separado.

Casos acumulados según nivel de gravedad (datos agregados Min. Sanidad)



Permite comparar la cantidad total según el nivel de cantidad en una fecha concreta.

Distribución de casos por sexo y edad según gravedad

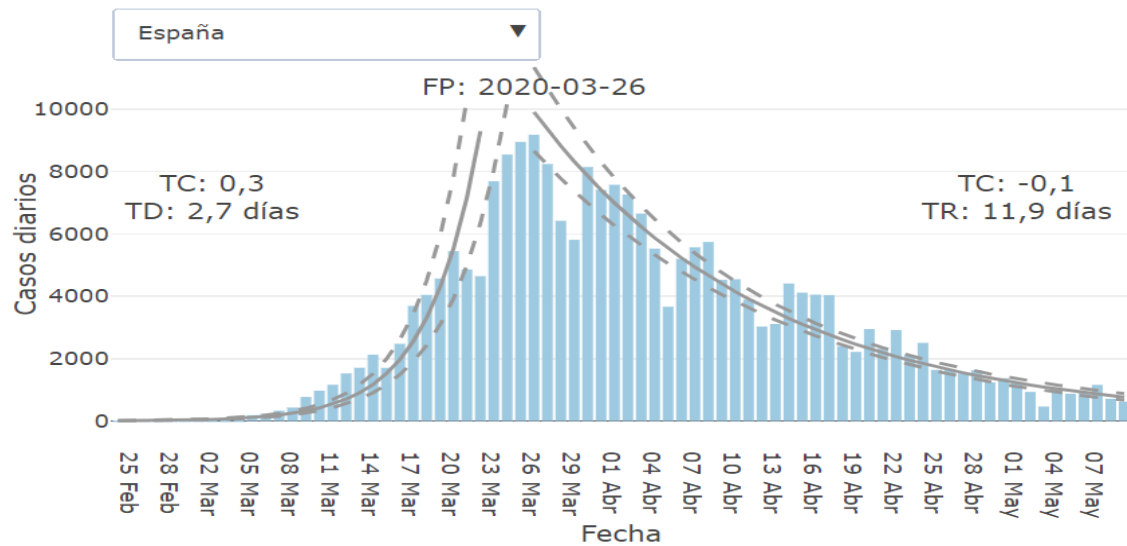


Muestra el porcentaje, según el nivel de gravedad seleccionado, sobre el total de casos haciendo una división en intervalos de cada 5 años y haciendo distinción sobre el sexo.

Por otro lado, nos enseñan datos de interés que resultan llamativos, como pueden ser:

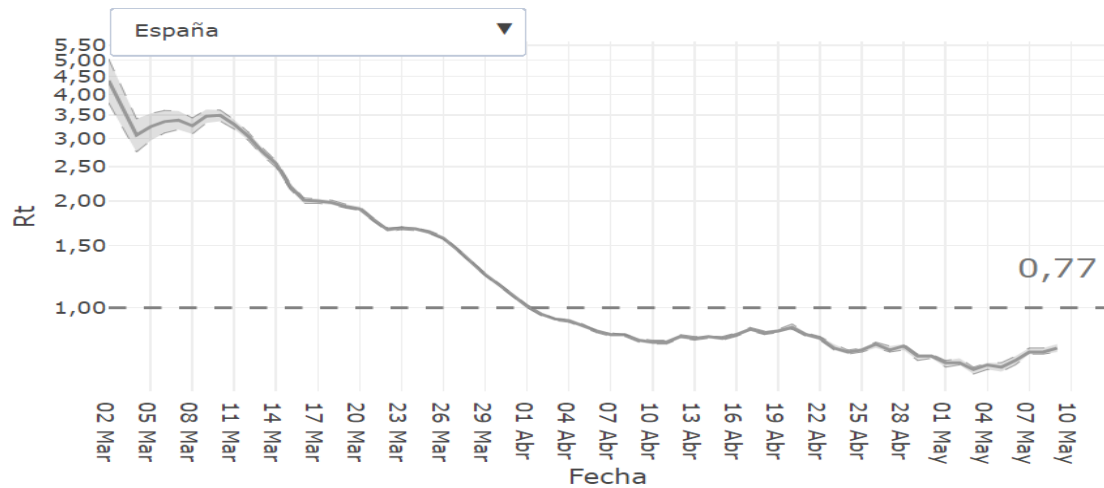
- Porcentaje de fallecidos con más de 70 años
- Porcentaje de fallecidos con al menos una enfermedad previa
- Proporción de sanitarios, distinguiendo entre hombres y mujeres, infectados.
- Porcentaje de sanitarios fallecidos, haciendo distinción por sexo.

Tasa de crecimiento



Curva de crecimiento de la pandemia. TC: tasa crecimiento, TD: tasa de duplicación, FP: fecha pico, TR: tasa reducción a la mitad.

Número reproductivo básico instantáneo (R_t)



Representa la evolución del número de reproducción básico instantáneo (R_t), que es el número promedio de casos secundarios que cada sujeto infectado puede llegar a infectar en una etapa de tiempo (t).

Por último, podemos descargar una base de datos actualizada cada día con las siguientes variables:

- CCAA
- Fecha
- Casos
- PCR+

- TestAc+
- Hospitalizados
- UCI
- Fallecidos
- Recuperados

En conclusión, podemos afirmar que la fuente que nos proporciona el gobierno ha mejorado desde el inicio de la pandemia. Esta mejora ha sido sustancial puesto que se partía de una pequeña tabla y se ha acabado mostrando diferentes gráficos con información más detallada. Por otro lado, la transparencia de datos continúa igual, puesto que sólo ofrecen una base de datos y con las mismas variables. Una propuesta de mejora sería añadir más de una base de datos, puesto que sería muy positivo tener datos abiertos de todas las variables a partir de las cuales se generan los gráficos o de las variables a partir de las que se muestran las cifras de mayor interés de la ciudadanía. Además, se podrían hacer públicos los datos individualizados ya que esto permitiría hacer nuevos análisis a las personas interesadas en manipularlos.

Los datos que aparecen en esta fuente oficial son proporcionados por las CCAA. Por tanto, sería interesante analizar la transparencia de alguna de las CCAA con el objetivo de comparar la calidad de datos que ofrecen y concluir si se encuentran diferencias entre ellas. La mejor opción sería comparar todas pero por motivos de limitación de tiempo lo haremos de unas pocas.

Comunidad Valenciana

Los datos de la Comunidad Valenciana se pueden encontrar en la siguiente página web:
[GVA/DadesObertes](#) .

A diferencia de los datos que ofrecía el gobierno, en la Comunidad Valenciana encontramos más de una base de datos en relación con la situación actual. En concreto 5 bases diferentes(contienen el enlace directo para descargarlas):

- Nombre de casos i taxa per cada 100.000 habitants per departaments de salut.
- Nombre de morts i taxa per cada 100.000 habitants per departaments de salut
- Total casos acumulats
- Percentatge i nombre de morts per rang d'edat i sexe
- Percentatge i nombre de casos per rang d'edat i sexe

Podemos concluir que en la Comunidad Valenciana tenemos suficiente información, renovada cada día, para poder seguir la evolución de la pandemia. No obstante, se podría mejorar subiendo bases de datos sobre hospitalizados, en UCI y recuperados en función del rango de edad y sexo, de la misma manera que con los fallecidos y los positivos. Además, subir los datos de manera individual permite sacar nuevos análisis y poder profundizar mejor. En conclusión, una posible mejora sería hacer públicos todos los datos recogidos y limpiados para que puedan ser reutilizados por cualquier persona. Esto ayudaría a favorecer otros posibles estudios a partir de los datos proporcionados.

País Vasco

En el País Vasco los datos se pueden encontrar en la siguiente página : [DatosAbiertosPV](#) .

Las bases de datos disponibles son(contienen el enlace):

- Pacientes hospitalizados (00:00 horas) (CSV)
- Positivos en los municipios de euskadi (08:00 horas) (CSV)
- Positivos en las zonas de salud de Euskadi (08:00 horas) (CSV)
- Fallecidos (CSV)
- Excel (EXCEL)

La página permite descargarse las bases de datos en cualquier formato.Suben un informe cada día con información más detallada que incluye otras variables como el número de fallecidos, recuperados, **número reproductivo básico**, etc.

En conclusión, el País Vasco presenta información de manera más detallada que la Comunidad Valenciana puesto que divide los positivos en municipios, zonas de salud y provincias. Además, la principal diferencia es que también aporta datos sobre los hospitalizados, en UCI y los recuperados por cada hospital.

Comunidad de Madrid

En la Comunidad de Madrid hemos visitado varias páginas puesto que en la única página web que hemos encontrado bases de datos sobre el COVID-19 es , [ComunidadMadridDatosAbiertos](#) , y en esta sólo encontramos dos bases de datos:

1. [Covid por Municipios y Distritos de Madrid](#)
2. [Zona básica de salud Covid-19](#)

En conclusión, podemos afirmar que la información respecto de los casos que dan positivo es muy buena. No obstante, no se ofrecen datos sobre los otros niveles de gravedad (hospitalizados, UCI, recuperados, fallecidos), ni según la edad, ni el sexo. Tampoco ofrece información en función de cada hospital. Sólo tenemos los datos según el municipio y zona de salud. Además, presenta casillas vacías en más de una columna.

Cabe añadir que en la siguiente página se pueden ver diferentes gráficos de la Comunidad de Madrid, [GráficosOficiales](#) .

Castilla y León

Se pueden encontrar los datos abiertos en la siguiente página web, [DatosAbiertos](#) .

Podemos encontrar hasta 15 bases de datos sobre el Covid-19 en Castilla y León. A continuación describiremos las variables de cada base de datos. Las bases son actualizadas cada día. Además hay una base en relación con los ERTes pero esta no la comentaremos puesto que hemos decidido no incluir bases de datos relacionadas con economía. Sí que añadimos la de contratos de emergencia puesto que muestra el dinero que se han gastado en el material necesario para cuidar a paciente y protegerse de la mejor manera del COVID-19. Bases:

1. [Hospitalizados](#)
2. [Situación personal sanitario](#)
3. [Tasa de enfermos por zonas básicas de salud](#)
4. [Tasa de mortalidad por zonas básicas de salud](#)
5. [Prevalencia coronavirus](#)
6. [Porcentaje de test realizados profesionales por categorías](#)
7. [Personas por lugar de fallecimiento](#)
8. [Enfermos en residencias por coronavirus](#)
9. [Mortalidad por tramos de edad y sexo](#)
10. [Enfermos detectados en atención primaria por tramos de edad y sexo](#)
11. [Enfermos detectados en atención primaria](#)
12. [Pruebas realizadas](#)
13. [Ocupación de camas en hospitales](#)
14. [Situación epidemiológica](#)
15. [Contratos de emergencia](#)

En conclusión, encontramos una gran diferencia a nivel de transparencia y calidad de los datos que proporciona Castilla y León a su portal de datos abiertos y las que hemos analizado anteriormente.

En resumen, podemos afirmar que Castilla y León es un modelo que se puede tener en cuenta para que otras comunidades mejoren su portal de datos abiertos. De esta manera, cualquier persona puede descargarse cualquier de las bases proporcionadas y manipularlas para hacer otros análisis que contribuyan al bien de la sociedad. Es una obviedad que si la calidad de los datos que se presentan es buena, las conclusiones que se puedan extrapolar a partir de ellos podrán ser de mayor utilidad y de esta manera aportar al objetivo común, servir a la comunidad y proporcionar conocimiento que pueda ser empleado con fines positivos. En la transparencia que presenta Castilla y León vemos que encontramos datos en referencia a todos los niveles de gravedad y que esta información es detallada de manera que se puede saber a qué hospital, provincia o municipio pertenece y dividida en grupos de edad y sexo. Agregan valor añadiendo las coordenadas que

facilitan poder hacer una representación gráfica mediante un mapa. Además, comparte datos de interés como puede ser la situación de los profesionales sanitarios o la evolución de la ocupación de las camas en los hospitales. Además, se hace público el dinero que se ha invertido en material para combatir la pandemia y podemos saber la cantidad de tests que se han realizado y cuántos de los que se han ido realizando diariamente han resultado ser casos positivos en Covid-19. Tener toda esta información aporta un gran valor a futuros proyectos, estudios o trabajos que puedan contribuir a proporcionar un servicio a la comunidad.

En este análisis podemos ver que la comunidad que aporta menos información es la de Madrid ya que sólo sube dos bases de datos con información de los casos que han dado positivo y además contiene errores. Según nuestro criterio, podemos encontrar una relación evidente y es que en este caso estamos hablando de la comunidad en la que más contagios se han producido. Esto requiere de una mayor labor del personal sanitario con los contagiados y a su vez son los sanitarios los que se encargan de recoger la información y los datos en relación con el virus. Esta situación de excesivo trabajo puede llevar a no poder notificar y registrar bien la información. No obstante, existen comunidades con un nivel de transparencia mucho menor que el de Castilla y León que no tienen las mismas cantidades de contagiados que la Comunidad de Madrid y que por tanto tienen margen de mejora en este aspecto. En conclusión, una manera de posible mejora en las comunidades a nivel de transparencia de datos puede ser observando como modelo a Castilla y León.

Por otro lado, el gobierno es el encargado de recoger toda la información que proviene de las comunidades y así nos muestra en su página web información al respecto de todos los niveles de gravedad por lo que han tenido que recibir estos de datos de todas las comunidades. Si este es el caso, todas las comunidades tienen la información respecto a todos estos niveles de gravedad de la enfermedad, pero no todos los están haciendo públicos. Hay que recalcar que la transparencia de estos sólo puede contribuir positivamente en la sociedad gracias a que amplía su posible manipulación a cualquier persona. Además, debe ser presentado y subido al portal de datos abiertos puesto que de esta manera cualquier ciudadano puede informarse de la evolución de la pandemia en su respectiva comunidad.

Además, gracias a la reunión realizada a través de Microsoft Teams (20/05/20) sobre el tema relacionado con los datos abiertos y la reutilización de datos, hemos podido realizar una pregunta a diferentes profesionales. La pregunta ha surgido gracias a la experiencia que hemos adquirido con el presente proyecto. De esta manera les hemos incentivado a que nos expliquen lo siguiente:

Pregunta:

-Sabiendo que los datos en relación con la pandemia son competencia de las comunidades autónomas y que cada comunidad está presentando bases de datos con diferentes variables, podemos concluir que algunas ofrecen más información y de más calidad. Pero, ¿son conscientes de esto y están tomando como modelo aquellas que mejor lo hacen para mejorar?

La pregunta ha sido respondida por tres profesionales:

Honorio Enrique Crespo Diaz-Alejo : Jefe de servicio. Subdirección de Transparencia (Ayuntamiento de Madrid).

Jose Ignacio Pastor Pérez: Cap del Servei de Transparència i Govern Obert en Ajuntament de València.

Victoria Anderica Caffarena: Directora del Proyecto de Transparencia del Ayuntamiento de Madrid. Asesora en materia de Gobierno Abierto del Grupo Municipal Más Madrid

En primer lugar, nos responde Enrique Crespo el cual cita:

‘No todas las comunidades aportan los mismos datos y hace falta las oficinas del dato para manejarlos correctamente. Se está viendo a nivel de España, de Europa y del mundo el descontrol absoluto de los datos que hay. Aun así en distintas comunidades se están manejando los datos bastante bien, empezando por Castilla y León y siguiendo otras que también ponen los datos a disposición y representaciones visuales. Aunque la coordinación como ocurre entre comunidades y administraciones muchas veces deja que desear.’

Posteriormente toma el turno de palabra Victoria Anderica:

‘Sobre el tema del Covid-19, en el portal europeo de datos abiertos hay muchos ejemplos de visualizaciones de distintos países y merece la pena mirarlo. Personalmente creo que en España no somos un ejemplo ni mucho menos. Hay comunidades que sí que lo están haciendo bien como Castilla y León, que realmente ha puesto todo su empeño en publicar los datos más relevantes y visualizarlos de manera que no sólo aportan un excel sino que permite que la gente lo puede entender mediante gráficos. Esto es importante porque los datos en una crisis como la que estamos viviendo son importantes para los analistas pero también para todos los ciudadanos. Creo que en España esto lo estamos haciendo mal a nivel estatal, de comunidades autónomas y a nivel local. Todos podemos aportar un granito de arena a dar contexto en estos momentos. Tenemos mucho que aprender.

La última intervención la proporciona Jose Ignacio Pastor:

‘Me gustaría comentar no como profesional, sino como ciudadano que la Coalición Pro Acceso pide al gobierno un mayor compromiso con el acceso a la información durante la crisis de la Covid-19. En esta coalición somos muchas entidades involucradas y planteamos una carta al ministerio. Destaco que los ayuntamientos, que formamos parte del estado, no hemos tenido todos los datos necesarios y con toda la protección necesaria porque esta situación no ha avisado y les ha pillado desprevenidos. Es decir, si una persona cuando entra a un hospital sabemos el dni, la tarjeta SIP, etc, con esos datos es muy importante cuando se está muriendo gente que los ayuntamientos inmediatamente a nivel epidemiológico supieran por ejemplo lo que ha pasado en un determinado barrio. Estamos hablando de elementos muy avanzados de contratos o contactos con las teleoperadoras que es muy importante, pero el sistema de conteo de uno a uno es tan fácil que se podría hacer. La pregunta sobre si se están falsificando los datos o no no se puede responder, habría que homogeneizar datos. Se están dando los datos por área de salud y la administración local, por comodidad, está siendo la

gran olvidada, pero el área de salud no tiene un componente político de decisión y nosotros por ejemplo como ayuntamiento si no tenemos datos no podemos actuar. Incluso es importante cruzar datos. Además estamos teniendo en cuenta constantemente lo que está pasando fuera en la sociedad y lo que le preocupa a la gente mediante los datos recogidos en encuestas.'

Formas de contagio del COVID-19 según la OMS

Respecto a la variable indique los métodos de propagación según las fuentes oficiales, proporcionaremos la información validada por la OMS.

Una persona puede contraer la COVID-19 por contacto con otra que esté infectada por el virus. La enfermedad se propaga principalmente de persona a persona a través de las gotículas que salen despedidas de la nariz o la boca de una persona infectada al toser, estornudar o hablar. Estas gotículas son relativamente pesadas, no llegan muy lejos y caen rápidamente al suelo. Una persona puede contraer la COVID-19 si inhala las gotículas procedentes de una persona infectada por el virus. Por eso es importante mantenerse al menos a un metro (3 pies) de distancia de los demás. Estas gotículas pueden caer sobre los objetos y superficies que rodean a la persona, como mesas, pomos y barandillas, de modo que otras personas pueden infectarse si tocan esos objetos o superficies y luego se tocan los ojos, la nariz o la boca. Por ello es importante lavarse las manos frecuentemente con agua y jabón o con un desinfectante a base de alcohol.

La OMS está estudiando las investigaciones en curso sobre las formas de propagación de la COVID-19. No obstante, hasta el momento esta es la información sobre cómo se propaga el Covid-19.

Esta información ha sido extraída de la página oficial de la Organización Mundial de la Salud (OMS). La página web es: [OMS](https://www.who.int/es).

Conclusión

Tras realizar un estudio exhaustivo sobre el nivel de transparencia de datos del gobierno central y autonómico en España, podemos concluir que la percepción de la ciudadanía tiende a valorar negativamente al central y esta valoración mejora con el autonómico. Gracias a la búsqueda de datos que hemos realizado, podemos comparar la opinión pública con la realidad.

En primer lugar, hay que entender que los datos sobre la pandemia son competencia de las Comunidades Autónomas y son estas las que transmiten los datos al gobierno central. A partir de esto, hemos observado que cada Comunidad ofrece diferentes bases de datos, con diferentes variables y con diferente nivel de calidad en los propios datos. Existen ciertas comunidades, como Castilla y León, que ofrecen una transparencia de datos sobresaliente y otras, como Madrid, que carecen de información y son presentadas con errores. Luego, el papel más complejo lo adquiere el gobierno central, encargado de unificar toda la información que pueda reunir de todas las comunidades. No obstante, la función principal la está realizando correctamente, es decir, representar los datos de manera visual para que toda la población pueda interpretar la evolución

del COVID-19. Éstos han sido actualizados diariamente hasta la fecha. Además, dejan a disposición una base de datos que reúne la información diaria separada por CCAA. Dentro de esta base se incluye información básica como CCAA, Fecha, Casos, PCR+, TestAc+, Hospitalizados, UCI, Fallecidos y Recuperados.

Si buscamos un mayor nivel de detalle, hemos de visitar los datos de las comunidades en la webs oficiales, de manera particular para cada una. Aún así, no en todas las comunidades, como hemos dicho antes, encontraremos un mayor nivel de detalle. En resumen, Castilla y León es un ejemplo a seguir por el resto de CCAA. Los expertos, con los que hemos podido comunicarnos, aportan que tenemos mucho que aprender en este tema, que la coordinación entre comunidades muchas veces deja que desear y que la administración local, por comodidad, está siendo la gran olvidada.

Por otro lado, podemos confirmar las hipótesis iniciales de que una gran parte de la población no contrasta la información a la que accede y que muchos de los gráficos que se publican en periódicos, televisiones, etc no son bien interpretados lo que también conduce a conclusiones erróneas. Lo hemos podido ratificar gracias a la inferencia. Ésta nos indica que entre 14,65 y 18,28 millones de personas en la sociedad española ha interpretado algún 'bulo' en referencia a los métodos de propagación como información válida. Se trata de un intervalo significativo que presenta el problema de la sociedad con la validez de la información que recibe. Además, entre 32.06 y 35.39 millones de personas en la sociedad española interpreta correctamente el gráfico. Partiendo de que España tiene entorno a 47 millones de personas y de que se trataba de un gráfico sencillo, consideramos alta la cantidad de personas que no saben interpretarlo. Añadiendo valor, sabemos con una confianza del 95% que no existe relación entre el sexo y no saber interpretar el gráfico, pero que sí que existe relación respecto a la edad.

Otra información importante que hemos logrado extraer a partir de los datos es que entre 13'58 y 17'08 millones de personas creen que el grupo de edad menos informado es el correspondiente a la tercera edad, sin embargo, este pensamiento es independiente al sexo y a la edad.

Para finalizar, ha quedado demostrado con esta situación excepcional el descontrol en cuanto a la calidad y organización de los datos se refiere. Teniendo en cuenta que los datos están catalogados como el futuro en todos los ámbitos, sería recomendable mejorar este aspecto a nivel nacional, si bien es cierto, como hemos visto en algunos análisis anteriores, que las personas más jóvenes tienen más conocimientos en estos aspectos, lo que favorece de cara a un futuro más o menos próximo.

Bibliografía

<https://opendatahandbook.org/guide/es/what-is-open-data/>

<https://okfn.org/>

<https://www.gob.mx/inafed/articulos/que-es-la-agenda-2030-para-el-desarrollo-sostenible>

<http://www.exteriores.gob.es/Portal/es/PoliticaExteriorCooperacion/Agenda2030/Paginas/Inicio.aspx>

<https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses>

https://dadesobertes.gva.es/va/dataset?q=&sort=views_recent+desc

<https://opendata.euskadi.eus/catalogo/-/evolucion-del-coronavirus-covid-19-en-euskadi/>

<https://www.comunidad.madrid/gobierno/datos-abiertos>

<https://www.comunidad.madrid/transparencia/situacion-global-del-covid-19-incidencia-municipio-y-distrito>

<https:// analisis.datosabiertos.jcyl.es/explore/?sort=modified&refine.keyword=COVID-19>

http://www.econ.upf.edu/~satorra/dades/moore_08.pdf

<https://cnecovid.isciii.es/covid19/>